June 1982 LBL-14713

Annual Report: FY 1981

RECEIVED
LAWRENCE
BERKELEY LABORATORY

NOV 16 1982

LIBRARY AND
DOCUMENTS SECTION

DCN/SEEDIS

The Distributed Computer Network (DCN) and Socio-Economic-Environmental Demographic Information System (SEEDIS)

Annual Report to the
Employment and Training Administration
U.S. Department of Labor

TWO-WEEK LOAN COPY

This is a Library Circulating Copy which may be borrowed for two weeks. For a personal retention copy, call Tech. Info. Division, Ext. 6782.

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

LBL-14713 c.2

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

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.

DCN/SEEDIS

The Distributed Computer Network (DCN)
and Socio-Economic Environmental
Demographic Information System (SEEDIS)

Annual Report to the
Employment and Training Administration
U.S. Department of Labor

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

This work was supported by the
Applied Mathematical Science Program
of the U.S. Department of Energy
under contract No. DE-AC03-76SF00098
and the Employment and Training Administration,
U.S. Department of Labor

Introduction

DCN/SEEDIS is a project which addresses both short term and long term computing needs of the Employment and Training Administration of the Department of Labor. Broadly speaking, the DCN addresses the short term and operational needs of ETA, while SEEDIS addresses longer term needs and directions. Both portions are based on the premise that fundamental changes are required in the way computing is done, if it is to respond adequately to user needs.

The purpose of DCN is (1) to provide a stable computing base, and (2) to provide a set of tools, conventions and user assistance which support problem solving directly by users, without requiring programmers as intermediaries. Software products (tools) of this quality and effectiveness have only recently become available. DCN endeavors to make these tools available, to instruct Regional office and National Office personnel in their use, and to create and provide new tools as their need becomes evident. The major tools currently used by ETA are: Datatrieve (a Digital product), the Software Tools (a product of past LBL research), the MSG system (a product of current LBL research), and CHART (a product of ongoing LBL research). Designed as modules which are independent, yet easily interfaced, these tools encourage end users to combine them as needed to solve data processing tasks. Command procedures and shell scripts provide the glue which holds these components together.

Over the longer term, SEEDIS responds to the needs of analysts, managers, and decision makers for broader, more accurate and more timely information. The requirements for achieving this lie primarily in the Data Management area. SEEDIS provides a testbed and a comprehensive set of operational facilities available now for addressing these information needs. Tools and techniques are provided for ***integrating*** data from diverse sources into a unified data storage system, ***locating*** information which is relevant to specific user needs, ***analyzing*** information to enable conclusions to be drawn and decision made, and ***displaying*** information for effective understanding of the data and conclusions. During 1981, SEEDIS was made significantly easier to use and was extended with new data bases. Substantial progress was made in the areas of file definition languages (part of metadata) and the incorporation of statistical analysis techniques.

Hardware Architecture and Basic Utilities

In Fiscal Year 1981 the Distributed Computer Network was composed of DEC VAX 11/780 machines at the ETA National Office, Seattle, San Francisco, and Lawrence Berkeley Laboratory. There are several necessary functions in supporting this network: keep the computers and basic software utilities running smoothly and reliably. In addition, LBL must monitor and evaluate current technology to maintain and enhance the cost effectiveness of DCN.

Support for the DCN is divided into three major areas: hardware and software support, LBL regional system management, and network management. The following sections discuss the LBL effort in providing these functions.

Hardware and Software Support

The development and demonstration of the experimental network requires several LBL personnel to support both the hardware and software systems. Support includes installation, test, and evaluation of software and hardware and the development of procedures which provide a reliable system for the users. Without adequate support, users quickly become frustrated and reluctant to use a system.

During Fiscal Year 1981 LBL continued to provide this hardware and software support for DCN. The hardware at each of the DOL nodes was expanded to accommodate the larger number of users and increased use at each site. Specifically:

(1) Additional Digital Equipment Corporation RP06 disk drives were installed at each DOL site. Seattle and the National Office now have three drives, with delivery of San Francisco's third drive awaiting expansion of their machine room. The extra online storage provided by these drives permits separating system and user files onto separate disks for gains in performance and reliability, and provides online access to large databases. Multiple disk drives also allow more versatile data backup strategies to be used.

(2) Additional memory was installed at each DOL site giving each machine three megabytes. Insufficient amounts of memory significantly degrade machine performance; with their current memory and disk configurations each DOL site should realize substantial performance gains.

During 1981 LBL installed Systems Industries disk and tape drives at LBL. These drives have been evaluated since January. Had they performed reliably, they could have provided DOL sites with increased storage at low cost. However, their performance has been quite erratic, and LBL recommends that only Digital drives be used at DOL sites.

The table shown on the next page illustrates the hardware status at the end of Fiscal Year 1981:

DCN Node Name	Disk Drives	Tape Drives	Memory Upgrade (Megabytes)
Region IX	2 RP06	TU45	From 1.25 to 3.0
Region X	3 RP06	TU45	From 1.00 to 3.0
ETADC	3 RP06	TU45	From 1.50 to 3.0
LBLG	1 RP06 1 SI300 Mbyte	TE16	4.0
LBLH	3 RP06 2 SI300 Mbyte 3 SI600 Mbyte (dual ported)	TU45 SI6250	4.0

The following major software items were installed or updated on DCN:

(1) A major new release of the VMS operating system was installed on all nodes. VMS Version 2 is a more mature product with improved integration of other software products such as DECnet. It has many new features and performance enhancements and has proved to be a significantly more reliable operating system.

(2) VAX-11 BASIC was installed on all nodes. This compiler is designed specifically for the VAX machine architecture rather than being an adaptation of a compiler designed for the smaller PDP-11 machine architecture. As such it eliminates many of the restrictions which were present in earlier versions and offers a substantial improvement in performance.

(3) VAX-11 COBOL was installed on all nodes. This version conforms to the COBOL-77 standard which provides increased functionality.

(4) Datatrieve-32, a new version of Datatrieve designed specifically for the VAX, was installed for testing at all sites. It removes many of the restrictions which were present with the older PDP-11 based version, particularly database size constraints. In addition to installation of the product release of Datatrieve, the DCN was invited to act as a field test site for future versions of the system; LBL accepted and field test versions were installed on all nodes. Such testing is beneficial to both parties participating in the test: ETA is benefited by having the opportunity to test new features in advance of the system's release. LBL's performance of tests on the field test system allows us to make recommendations on needed improvements or changes. Digital benefits from the field testing in that its systems are thoroughly tested before release, and is given feedback on the product's usability.

The DRS data management system was installed and tested at LBL.

The table shown on the next page shows the status of software on DCN at the end of Fiscal Year 1981:

Software Package	DCN Node			
	LBL	RIX	RX	ETADC
VMS	2.3	2.3	2.3	2.3
DECnet	2.0	2.0	2.0	2.0
Fortran	2.3	2.3	2.3	2.3
Datatrieve	2.1	2.1	2.1	2.1
Cobol (native mode)	1.2	1.2	1.2	1.2
Basic (native mode)	1.2	1.2	1.2	1.2
Basic Plus 2	1.6	1.6	1.6	1.6

LBL Regional System Management

LBL continued to provide support for the DOL regional system managers. Specific instances of this support by way of diagnosing hardware and software problems were documented in detail in Fiscal Year 1981's monthly reported.

In addition, procedures were developed at LBL to facilitate operation of the regional systems and consultation was provided on the more difficult hardware and software problems. Programs were developed to gather machine statistics to better quantify system performance. LBL staff continues to modify system parameters on remote nodes to optimize system performance in response to increased user load and enhanced hardware configurations.

Network Management

The network management function is required to keep the communication between all the machines running efficiently. Equipment must be tested and monitored for reliability.

1981 marked the first year of full production network operation. The network was no longer a test entity, but was functioning as a reliable, well utilized tool. The Paradyne modems were unable to pass acceptance tests and were replaced by Bell modems which proved quite reliable after overcoming initial installation difficulties.

LBL and Region 9 were test sites for DECnet Phase III. This version of DECnet provides remote terminal support, automatic route-through, and enhanced file access. Remote terminals allow users to login to a remote site using the DECnet communication links, eliminating the need for separate modems. Aside from a reduction in transmission speed, the presence of an intermediate host machine is transparent to the user; the user appears to be directly logged onto the remote node. Automatic route-through makes each remote

node appear identical to the user. Each remote node appears to be logically adjacent to the local node, although physically there may be several intermediate nodes. This simplifies network access for the user as it eliminates the need to have knowledge of the physical network topology. The enhanced file access provided by Phase III allows all types of files to be accessed via the network eliminating a previous restriction. Region 10 and the National Office were updated to DECnet Phase III when it completed testing and became an officially released Digital product.

The following two tables show the layout of the network and modems installed at each DCN node at the end of the fiscal year.

Network Configuration			Telephone Modems	
DCN Node	Node(s) Connected to	Type of Link	Number of Lines	Number Added This Year
LBLG	LBLH LBLJ	1 Megabit 250Kbit	5	1
LBLH	LBLG RIX DGATE	1 Megabit 9600 bps 250Kbit	6	2
RIX	LBLH RX ETADC	9600 bps 4800 bps 4800 bps	8	4
RX	RIX	4800 bps	5	5
ETADC	RIX	4800 bps	12	10

Data Base Management Systems

Distributed Data Base Management

A crucial difference between the DCN environment and the classical data processing environment is the electronic connection of geographically separate computers into a unified network wherein resources may be shared throughout ETA. In particular, data may be distributed and managed at local ETA Regional Offices under Regional control, while at the same time, this data may be accessed and shared both between Regions and by the National Office of ETA. Thus, with the proper computer environment, the optimum mix of local processing and management, combined with full satisfaction of central administrative requirements, may be achieved.

In order to effectively utilize this networked base of computing machinery for the management of data, research into distributed data base management is vital. This research leads to a better understanding of those processes which are fundamental to efficient and fail-safe operation in a distributed environment. The components of distributed data management include aspects of:

file allocation,
caching,
efficient network utilization,
access to "foreign" data bases.

File allocation is the process of determining the permanent location of data within the network. *Caching* is the allocation of temporary space for high use files and the control of space allocation. *Efficient network communication* depends on good protocols for query descriptions, and the correct division of labor between local and remote processing for minimization of computing and communications costs. *Access to "foreign" data bases* allows one system to access data maintained by another system, and will be discussed in the following section.

Since the entirety of SEEDIS data bases will not be resident on a particular DCN node, network access to non-resident data must be provided. For other data bases which are distributed according to control and function (such as the RAS system being distributed to regional ETA offices) distributed access must be provided for National Office needs.

During FY 1981 areas of investigation in distributed access and control for a homogeneous network included:

file sharing capability
the ability for programs and users to access remote files;

device sharing capability
the ability to access remote peripherals;

program sharing
the ability to transmit a system to a remote computer and have it run there;

retrieval request processing
the tradeoff between remote data access for retrieval of specified data and access of complete files for pro-

cessing at the host node;

data manipulation decisions
regarding remote or local processing;

recovery and retransmission
the ability to "hold" retrieval requests if remote machines or peripheral devices holding data are off-line.

SEEDIS implemented network retrieval for nonresident data using task-to-task communication within the DECnet facility, so that data location was transparent to end users, except for access time. In addition, whenever a SEEDIS session is initiated, it automatically initiates a process which polls the nodes and gathers information on the current location of various SEEDIS disk packs and data bases.

DBMS Selection

While the ongoing development of DBMS systems demonstrates that the best system-of-choice in the future may not be the best one currently available, critical project requirements dictated procurement of a system upon which development could begin. On the basis of this evaluation, two copies of the DRS system marketed by Advanced Data Management of Princeton, New Jersey were acquired and installed at LBL and the National Office.

Classes in DRS provided a more detailed picture of its strengths and weaknesses, as did participation in the first DRS Users Conference, which was held in Chicago in June 1981. An 18 page report of DRS deficiencies was presented to the vendor.

The majority of ETA DCN users are involved in the production of applications developed under Datatrieve. Some of these applications are now over three years old. Datatrieve is a data manipulation system written by Equipment Corporation to facilitate the entry, updating, deletion, and validation of data. The system employs an English-like query language to specify data retrieval and ad-hoc report generation, as well as a data definition mechanism for specifying the format in which the data is stored in the computer.

Digital has invested significant effort to make the system as flexible and comprehensible as possible for those users at whom the product is aimed: end users, namely managers and analysts rather than data processing professionals.

At the end of 1981 Digital announced the release of the VAX 11/780 native mode version of the system with increased capabilities. Primary among these new capabilities, is the concept of remote domains. With VAX-11 Datatrieve it is now possible to define data bases on local nodes with the actual data resident on distant nodes. The actual location of the data is transparent to the end user during data access.

Performance criteria were developed and performance testing was begun on DRS, Oracle, Ingres, and Datatrieve.

LBL staff continued to obtain information about data base management systems as new ones became available and existing ones changed.

Data Base Management

Data management, broadly interpreted, encompasses all those mechanisms which connect users and application programs to data they require. Data management is especially important to ETA because it can help to reduce costs by facilitating greater sharing of data and data processing tools. Data sharing depends on an adequate theory and implementation of underlying data description concepts, as well as upon *standards* for communications between applications modules, data description information, and underlying data management software. Data description and communications standards in turn provide a convenient (and necessary) starting point for the development of distributed database management systems, and systems capable of using data from heterogeneous underlying hardware and software systems.

Metadata Research

When there are large numbers of databases and individual data items, it becomes difficult, if not impossible, for users to remember what data is available or what a particular data set is called. This is especially true for ETA, where there are many large and diverse databases (e.g., census, ESARS) and many users who are not particularly familiar with either computers or the wide variety of available data. Such situations call for databases of information about data, called metadata, and tools to manage such metadata.

Metadata management is a relatively new area of computer science research, and many problems remain to be solved. Ideally, one should be able to create and manipulate metadata using the same tools as one uses for the data itself, but there are few, if any, systems that provide even rudimentary integrated capabilities of this sort.

Three aspects of metadata that are particularly important for ETA and the DCN/SEEDIS Project are:

data definition language (ddl);
data dictionaries;
data indexes.

Data definition language is a nonprocedural means to describe various aspects of data, such as physical characteristics, logical structure, statistical properties, labeling information, reference documentation, security provisions, etc. at the database, data record, data item, and individual item instance levels. The ddl is used initially to construct data definition files or schemas, but it can also be used to connect applications programs to databases, to generate data dictionaries and indexes, and to simplify data manipulation operations by providing information and checking facilities that would otherwise have to be built into the manipulation procedures.

Since user needs are always changing, and new needs may call for new types of metadata information, one very important consideration is how to make the ddl sufficiently flexible so that database administrators can add new types of metadata information and revise logical structures from time to time without having to rewrite applications software or reload substantial portions of either data or metadata. The language must also be able to unambiguously define a variety of data structures and to pro-

vide a means of translating between different structures. At the same time, the ddl needs to be simple enough so that non-programmers can understand and use it.

During the past year, LBL staff have developed specifications for an enhanced version of the SEEDIS/Codata data definition language, which will provide greater flexibility and increased ability to handle some of the more complex data structures that ETA wishes to make easily accessible, including multi-dimensional tables (e.g., 1980 Census data), hierarchical records (e.g., Current Population Series, County Business Patterns), and time series (e.g., Bureau of Labor Statistics LABSTAT data).

Data dictionaries or directories are the central repositories of metadata information. Central data dictionaries help isolate application programs from details of specific databases and from changes in individual database formats, so that applications can be applied more generally to a wide variety of data. Dictionaries are important not only to control and manage data but also to reduce costly redundant storage of data and metadata. In addition, they provide a visualization of database structure and contents for various kinds of users.

During the past year, LBL staff have been investigating methods for automatic generation of data dictionaries from data definition files, and we have been experimenting with alternative formats which can help make data easier to find and understand. For example, the current SEEDIS format for 1980 Census data dictionaries represents a substantial improvement over documentation distributed by the Census Bureau itself, and we hope that it may become a future standard for similar types of documentation.

Metadata indexes reference databases, data items, entities to which data pertains. Such indexes help facilitate both ad hoc queries and the development of new applications based on existing files. During the past year DCN/SEEDIS and related projects at LBL have studied alternative ways to automatically index information from large statistical databases and provide users with indexes and thesaurus-like facilities for browsing the contents of an entire data collection -- accross many different databases, entities, and data items. The SUBJECT system developed by a related project provides underlying capabilities that may make it possible to implement a variety of indexes and thesauri within SEEDIS, so that users could browse metadata information by subject, time period, or simply key word, as well as by database and geographic level.

Standards for DBMS Functions

Standards for DBMS functions are required to reduce the complexity of interfaces between application programs, metadata, and data access systems. With a standard as a common point of departure, each component in a distributed system need only conform to one interface, rather than a multiplicity of interfaces, one for each other component. Moreover, standards facilitate the re-use of modules from one application to

another and the substitution of one module of a particular type for another with little impact on other components in the system. This building block approach results in substantial savings in the effort required to implement ETA projects.

Standard interfaces can be developed at several levels, including self-describing data interchange formats, standardized access routines for data and metadata, standard data manipulation tools.

The current SEEDIS CODATA file format has been a very successful self-describing interchange file format for several years, while the codata tools have provided standardized access and manipulation routines which operate simultaneously on data and metadata. By standardizing the file format and data access interface, access level modules could be written to specifications without regard for the details of application programs which used them. In turn, new applications were based on CODATA access routines and could be implemented much more quickly.

However, there are now a number of compelling reasons to substantially revise the CODATA interface specification to provide:

increased functional support for different applications;
a wider variety of physical file access methods;
facilities to handle complex data structures (such as LABSTAT);
compatibility with DBMS systems, such as will be used for RAS;
improved network capabilities for distributed DBMS functions.

During the past year, DCN/SEEDIS began work on an enhanced data interchange format and specifications for standardized access routines for data and metadata in order to meet these ETA needs.

Data input, output, and manipulation is another area where standardization can save resources in a variety of areas, including: initial design and programming effort, ongoing program maintenance, quality of procedures.

At present, most applications programs and even DBMS facilities require a substantial amount of custom input, output, and manipulation routines that do essentially the same job. This results in wasteful duplication of effort, higher maintenance costs, and lower quality procedures.

Examples of the many different types of operations that could be put into a standardized form might include conversion of dates to a single internal form, checking a value against a list of valid codes, or putting out decimal numbers with preceding dollar signs. Following the software tools principles developed by colleagues at LBL, it should be possible to develop a library of small, discrete routines, each specific to one particular type of task. As in other software tools, each routine should read a "standard input" and write a "standard output" so they can be chained together to do a number of input validation and transformation steps on a single data value or stream of data values. These standard data manipulation procedures could be included as part of a data definition language and thus part of the central data definition rather than individual routines.

Access to Data Bases on Non-DCN Machines

Automatic access to databases on non-DCN machines will assume increasing importance in the coming years as ETA analysts need to integrate data maintained by a variety of data management systems

on different machines with their own working databases on DCN machines. Distributed access will be required because:

ETA sources (SESAs, Prime Sponsors, BLS) maintain data on non-DCN machines;

originating organizations should be responsible for their own data;

each system is optimized for different functionality;

long-term duplicate storage of data is inadvisable because of difficulties in maintaining consisteny as well as storage costs.

Access to heterogeneous machines in a transparent manner is a research problem of significant magnitude. The normal modes of communication to remote machines include: integrated network connection (such as the ARPANET); remote job entry (simulated card-input and printer-output); remote terminal connection (simulated human-input and screen-output).

All of these approaches require a considerable investment in hardware and system software. For both hardwired lines and dial-up access, a database of remote machines and their database characteristics is required. The structure and utilization of such network metadata databases is a substantial research problem.

The DCN/SEEDIS project and related projects within LBL have made a variety of contributions toward understanding and solving these problems during the past year. Work on virtual machine architecture provides a basis for databases distributed across heterogeneous systems. More powerful and flexible data definition language constructs necessary for representing remote hardware and database characteristics, as described above, have been identified and are being tested. Experiments in accessing foreign databases such as LABSTAT have led to a better understanding of data translation techniques and to the consideration of database structure translation, e.g., how to represent a foreign database structure in a local data dictionary and transform that automatically to another working database on the local machine. Finally, initial specifications have been drafted for a local integrated network connection to the LBL Computer Center's Tape Robot and CDC computers, which will permit interactive access to the entire archive of 1970 and 1980 Census data down to the block group and ennumeration district level.

Graphic Display Systems

CHART Development:

There has been considerable interest recently in new graphic forms for analysis and data display. One such form - continuous tone color shading - has been incorporated into CHART to address issues of data quality and decision making commonly encountered in energy- and labor-related data. Data such as estimates, projections, or proxy variables are often imprecise and uncertain, and the decision making process not well defined. A shift in analysis from the numerical values themselves to broad category terms appropriate for each data variable, such as "high" or "low", helps diminish the misleading effects of undue precision. Another motivation for a shift in viewpoint is the opportunity for cognitive economy. Attention can be directed where judgment is most needed, and unnecessary distinctions suppressed where the issue is clear cut.

This is achieved by a model for category terms based on fuzzy set theory, and a continuous tone color display to show where and how well data fit the category description.

Two application areas are being investigated. Within regional offices of the Department of Labor, these displays have been used for monitoring and evaluating prime sponsor performance, as well as for routine management information reporting.

Similar displays are proposed for a geothermal energy application. An important federal policy promotes private sector exploration and development of geothermal sites on federal lands. Leasing of federal lands to conduct exploration activities is a first step which makes available identified economic resources for potential development. The displays present quantitative indicators of development potential to analysts and planners, and are designed to respond to questions such as "which geothermal sites have the best potential for successful development?" or "which sites suffer from environmental delay but otherwise have good potential?".

CHART Maintenance:

The effective table size has been expanded to use all available virtual memory. The maximum table size is a function of the product of the two dimensions so that, for example, narrow tables can be longer than wide ones. Space is sufficient for up to 7000 rows by 100 columns, for example.

A software character font has been included. Characters in the this font can be drawn in a continuous range of sizes, use a minimal number of strokes to speed plotting time, and present a clear crisp appearance. Work was begun on an outline and solid fill Helvetica font especially suit-

able for large size characters, such as main titles and labels.

Default page sizes for non-graphic terminals were set for the 80 character ADM3 terminal rather than the DEC-writer or line printer.

Data files can be written from CHART in standard Codata format. Existing map keys can be preserved or new ones entered.

Title lines can be positioned anywhere on the page using either graphic input, or explicit page coordinates.

Mapfiles:

Software for generating and maintaining cartographic base files for CARTE was enhanced. A unique feature of Seedis is the large number of geographic levels supported. The great majority of these, such as SMSAs, are defined in terms of counties. Software was developed to

Prepare a mapfile for aggregation by inserting proper geocodes;

Perform the aggregation;

Subtract island areas from surrounding polygons;

Instrument the aggregation process to verify correctness;

Conveniently fix any polygons in error.

Several new mapfiles were generated in this way:

New England Consolidated Metropolitan Areas (NECMA) as of 1977 and 1979, Bureau of Economic Analysis areas for 1977 intersected with 1977 NECMAs.

Other functions added to the repetroire of map manipulation tools are: placing polygon labels, scaling, splitting, and joining polygons, and systematic replacement of selected polygons. These functions were used to construct a U.S. by county mapfile complete and consistent with both 1970 and 1980 county definitions. In Virginia, some 35 independent cities designated as counties are represented schematically by small squares inside a larger, more conventional, county. One small county in Hawaii is also treated this way. In a few other cases, counties are broken into fragments, where each fragment may have different 1980 codes, say, but a common 1970 code (or vice versa). The transfer of tract mapfiles to the VAX was completed and all files were installed in Seedis. Census tracts for approximately 250 SMSAs are available. Congressional District boundaries (96th Congress) were obtained from the Census Bureau, converted to longitude/latitude, and installed in Seedis. This file appears to have been derived from the original U.S. by county file, so state and county boundaries will match if these are overlaid. Canadian census division boundary files were processed into nickel format and installed. As a by-product, the great lakes boundaries were extracted and can be combined with the U.S. boundary. A world map by nation from World Data Bank I was installed with nation codes according to 1980 FIPS definitions.

One way to represent dataflow diagrams or information networks is in terms of maps. Using mapfile descriptors to set Grafpac attributes

on a polygon by polygon basis, sample diagrams were composed to illustrate reports and test the feasibility of using this approach.

Sample frames of color typography were generated in this way to investigate user interface design issues (see color Xerox pages).

User Documentation

Interfaces

During FY1981 significant progress was made in redesigning the user-machine interface of Seedis. The goal of this research was to determine conventions and explore visual/verbal standards which would improve the appearance of different modules within Seedis by making them communicate clearly and consistently. Considerable effort was expended to accomplish the following:

(1) Establish a spatial grid for all typographic elements of frames of information displayed on the screen.

(2) Establish system-wide typographic conventions for prompts, lists, module identification, and other documentation devices.

(3) Establish some naming and terminology conventions, especially with respect to global commands available throughout the system.

(4) Rewrite and, where necessary, expand the on-line help message for the upper levels of Seedis. In particular, each message was reformulated to be consistent in vocabulary and logically independent of other frames of information.

(5) Re-design the user interface of a new module, *QUERY*, being introduced into Seedis. This constituted a test situation to determine whether the proposed conventions for Seedis could be used for new modules with enhanced functionality.

(6) The proposed frame graphic design standards have also been adapted for use in printed documents, especially the Seedis release notes.

(7) The Seedis graphic design conventions have begun to affect the way in which data directories appear in Seedis, both online and offline. During Fiscal Year 1982 major efforts will be made to adapt the general Seedis user-oriented graphic design standards to the eye-readable format of data directories.

Documentation

Command files were implemented to automatically update the SEEDIS Release Notes and accompanying on-line documentation, each time a new data base is installed in SEEDIS. Internal documents on procedures for installing data were revised and expanded.

Command files were implemented to process data dictionaries for either (a) printed documentation, 62 lines per page, or (b) on-line documentation, 21 lines per page. Only a single source document needs to be maintained, and a correct page-numbered table of contents is automatically produced in each case.

Command files were implemented to automatically print, on the LBL computer center's narrow-paper printer, as many copies as requested of a given SEEDIS document. Over a dozen documents are available. During FY1982 this facility will be extended to cover the 100-odd data dictionaries, and installed in SEEDIS with on-line documentation.

Command files were implemented to automatically archive and catalog machine-readable logs of every SEEDIS run. So far, over 13,000 logs have been permanently archived for future statistical analysis. Four times per year, a special command file is run to automatically summarize SEEDIS usage statistics for the previous quarter.

Statistical Analysis

Aggregation and Disaggregation. Additional geographic levels, especially those required to describe 1980 Census geography, were incorporated in Seedis. Among others, these include the following:

Level	Description
COUNTY80	Counties, 1980 Census definition
CNTY7080	1970/1980 Census county pieces
SMSA81	Standard Metropolitan Statistical Areas, 1981 definition
NECMA81	New England County Metropolitan Areas, 1981 definition
PLACE80	Places, 1980 Census definition
MCD80	Minor Civil Divisions, 1980 Census definition
TRACT80	Tracts, 1980 Census definition
EDBG80	Enumeration and Block Groups, 1980 Census definition
EDBG70	Enumeration and Block Groups, 1970 Census definition

Enhancements were made to geocode correspondence files, to permit automatic aggregation and disaggregation to and from these new geographic levels. A file was purchased from the Census Bureau, which will permit conversion of 1980 Census data to 1970 Census tract boundaries, for display with 1970 tract map files. The method of automatic aggregation and disaggregation was described in a paper presented at the Workshop on Statistical Database Managment, at Menlo Park, California on December 2-4, 1981. The method was applied to the PL94 Reapportionment data installed in Seedis, to permit data to be obtained for either 1970 or 1980 Census counties or for 1970 or 1980 county pieces. Files were constructed, based on newly published FIPS definitions, which will allow the selection of states or provinces for any selected nation(s) of the world; the first applications will be in the analysis of Canadian acid rain and air quality data.

Model. An experimental version of a new "Model" module was installed in Seedis. With this module, the user can estimate air quality for any geographic entity whose location (in latitude and longitude coordinates) is known. Input data for the model are air quality measurements at specific monitoring stations, obtained from the SAROAD (Storage and Retrieval of Aerometric Data) system maintained by the Environmental Protection Agency. Input parameters for the model are specified by the user, in a manner analogous to the selection of data elements from archived data bases. In the future, similar models can be implemented to estimate (for example) population density as a function of latitude and longitude.

Codata Tools. Enhancements to the Codata Tools include problem fixing and the implementation of two new tools: Coroagg and Ddffmt. Coroagg, a generalized version of Coroadd, permits aggregation of data for user-specified combinations (sums or differences) of geographic areas. Fractional combinations (for example tract 1 minus 30 percent of of tract 2) are allowed. Ddffmt produces a one-line Fortran format statement corresponding to a given Codata file; this format statement can then be furnished to a statistical package or a user-written program for automatic processing of the Codata file. Experimental versions of the Codata tools have been written to provide extensions to the present file descriptions. Major enhancements to the present Codata file format, for future implementation, are under discussion.

Command files were implemented to permit the Codata Tools to be completely reconstructed, with user-specified dimensions, at any node on DCN.

Statistical Packages. Experimental interfaces between the Codata file format and two statistical packages (SPSS and MINITAB) have been implemented, and used in some applications. Problems with these interfaces have pointed up the need for a coordinated solution to the general problem of data format conversion. Long range plans are being discussed in detail.

Query. The user interface to Query was greatly improved and now stands as a standard example for future modifications elsewhere in Seedis. An important new command, permitting the automatic addition of a series of data elements, was implemented. Command files have been implemented which permit Query, (including Langpak, its principal subsystem) to be completely reconstructed from source code.

Newdata. Newdata was enhanced to make it run three times faster than previously. Command files were implemented to permit Newdata to be completely reconstructed from source code.

Area. The Area module was enhanced to permit selection of 1980 Census subcounty geographic entities.

Log Files. User entries are uniformly echoed to archived log files, which can be used to run SEEDIS in a batch mode. This facility has proved extremely useful in accomplishing lengthy or repetitive tasks.

1980 Census Project

The 1980 Census Project provides integrated data management within the SEEDIS information system for the estimated 50 billion characters of machine-readable data from the 1980 decennial census. The project expects to produce about 4 million pages (printed and on computer outputted microfiche) of reports tailored to ETA planning needs for a wide variety of geographic entities including states, counties, SMSAs, Federal Regions, Prime Sponsors, cities, and census tracts. Reports are being specified by an ETA Census Planning Committee in cooperation with LBL technical staff and will augment Census Bureau publications by providing unpublished data from machine readable files, computed indicators tailored to planning requirements for ETA programs, and integrated computer graphics (maps and charts).

Executive Summary

During Fiscal Year 1981 significant progress was made in developing an environment for the processing of 1980 census data, formatting and laying out of reports, creating a generalized report generation system for producing the reports, and testing aspects of the system on preliminary released data from the Census Bureau (including the PL-94 "Reapportionment Data"). The major accomplishments of the Census Project in the past year have been:

- LBL has, after considerable consultation with DOL analysts, designed a standard typography and layout for the Census reports which conforms to industry standard principles for graphic design.

- LBL has completed a thorough investigation of computerized hardware and software for automatic typesetting and printing of the Census Reports. This investigation will result in the use of the Xerox-9700 laser printer for production of the Census Reports, as well as additional copies of the reports to be produced on microfiche.

- LBL has created a generalized report generator for automatic formatting of the report specifications by non-computer analysts or computer technicians. This report generator will provide substantial cost savings in report creation and production processing over methodologies used in the 1970 Census Manpower Reports.

- LBL has modified the generalized SEEDIS software to process data from the 1980 census, and thus provide for long-term, ongoing information requests, as well as Census report production. The basic areas of enhancement were:

Significant design progress was made in enhancing the SEEDIS data model to incorporate those changes necessary for effective processing of the 1980 Census data files.

Local network connections between the DCN VAX machines and the LBL computer center machines were installed and made operational. This opens an access path to the Automatic Tape Library mass storage

tape robot, which is currently the only viable device for archiving the 1980 Census.

• LBL designed a production-tracking system to keep track of tapes, reports, addressees of recipients, and production and processing using the Datatrieve data storage and retrieval system.

• LBL has established and priced the components of production and distribution of the final printed reports through the National Technical Information Service (NTIS) of the Department of Commerce.

Project Components

Specific components of the Census Project are:

(1) Design and development of report table specifications in cooperation with Region IX and National ETA task forces

(2) Investigation of hardware and software for high quality typography and graphics

(3) Acquisition and installation of hardware sufficient to support 1980 Census processing

(4) Software development for report definition and automated report production

(5) Preparation of training materials for actual report definition input by Region IX ETA personnel

(6) Software development for report production scheduling and distribution

(7) SEEDIS system enhancement to provide an adequate environment for Census processing

(8) Acquisition and installation into SEEDIS of summary tape data and geographic base files from the 1980 decennial census

(9) Report production

(10) Ongoing information requests.

The remainder of this report addresses each of these component areas in more detail.

1. Report Table Design and Specification

The 1980 Census report tables, including tabulations and computations have required careful design. The design includes general visual layout and precise typographic specifications that will be created in cooperation with Region IX and National Office ETA task forces on the 1980 Census. Draft documents (tables, texts, charts, and maps) were revised for legibility, comprehensibility, and visual appeal, as well as enhanced to take advantage of typesetting technology on an ongoing basis throughout the Fiscal Year. These tables are being supplemented by textual pages of explanation and special purpose graphics such as charts and maps. Within design meetings with the ETA Working Group on Census Report Design and LBL technical staff, LBL has provided input on organization of the contents of the reports as well as the visual layout and specification of the pages.

In structuring the reports, LBL has clarified the contents of table stubs and headers and standardized titling within and among different reports. In addition, LBL proposed an index style and some explanatory text portions of prototypical reports. LBL also

catalyzed a tentative selection of specific items from each of the reports that would be visually highlighted through charts and maps.

With regard to visual layout and specifications, LBL proposed detailed examples of typical pages showing tables, texts, charts, and maps. In addition LBL has developed a precise, thorough, but still tentative typographic specification for all the tabular pages. This typographic specification lists in prose style the explicit typographic distinctions which characterize particular elements of any table, e.g. the size, position, boldness and font of second-order stubs.

Reports that have been graphically designed will most clearly and effectively demonstrate the content of census values which are of interest to ETA planners and project monitors (Federal Representatives). A compendium of table designs, general report layout, and typographic style specification has been prepared and will be distributed upon request.

2. Investigation of Typesetting and Graphics Hardware and Software

An investigation of the state of the art in computer driven high speed typesetting printers and graphic output devices is essential to the efficient production of a large volume of reports of high visual quality. We have focused on the capabilities of the Xerox 9700 laser printing device which prints at 10,000 lines per minute and allows interleaved graphic output. Of particular interest are the typesetting software (which includes multiple typeface fonts and forms rules) and graphics driver abilities (which may have to be written at LBL).

A thorough investigation has been made of hardware and software for computer-driven high speed typesetting printers and graphic output devices, with particular emphasis on the Xerox 9700. We have made a number of special test runs using a device run by a print shop in San Francisco. Software to run the printer remains a problem because a complete range of such software is not generally provided by the manufacturer. Local installations often develop non-standard implementations for even conventional typographic tabular and text settings. Software for charts and choropleth maps does not yet exist commercially. We have made contact with experienced consultants who develop software for this machine and are pursuing negotiations and further study with other consultants who offer software development as well as production services.

3. Acquisition and Installation of Additional Hardware

Additional disk and other hardware proposed as necessary to handle 1980 Census processing was not funded during Fiscal Year 1981. It is expected that such hardware will be installed during Fiscal Year 1982.

4. 1980 Census Report Generation

For the 1980 Census report preparation to proceed in a cost effective manner, it was decided that the methodology utilized for the 1970 Census to develop the Manpower Indicator and other packages (all hand-coded in FORTRAN) could be substantially improved in 1980. Our approach has been to design and implement a report generation tool sufficiently general to produce reports in at least as wide a variety of formats as were available for 1970 Manpower Packages, but which may be directly designed and input by DOL professional staff with minimal assistance from LBL computer scientists, allowing the latter to concentrate on problems of processing, efficiency, and delivery.

Functional Characteristics Desired in the Report Generator:

As presently operational in a working prototype, the 1980 Census report generator includes the following capabilities:

- Left and right justification and centering
- Line (row) positioning
- Tabsets and Tabs
- Computation
- Horizontal and Vertical percentages
- Totals and Subtotals
- Page headers and footers
- Underlining and boldface
- Automatic handling of missing data flags
- Intermixed graphics and text (i.e. charts and maps on the same page as some tabular reports).

All of the above mentioned features, excepting the last two have been implemented in a working report generator, during Fiscal Year 1981. This report generator has been tested on the 1980 Census Richmond Test STF-1A file and P.L. 94-171 reapportionment files supplied by the Census Bureau, and other SEEDIS data files (the 1974 Census of Agriculture, for example).

5. Training and Report Specification Entry by DOL

Training materials will be prepared to enable DOL analysts and their clerical assistants to input specifications for the 200 pages of tables for the 1980 Census Reports directly into the computer on their own DCN node. During Fiscal Year 1981, two preliminary drafts of a users guide for report input were prepared.

6. Report Production Scheduling and Distribution

The actual Summary Tape Files (STF) of the 1980 Census will be released in a phased fashion over the next few years. Within each STF, data summarized for smaller states will be released first, then larger/states, by population size. This will create a need for scheduling software to manage STF installation, compression, archiving, report production on a state-by-state basis, as well as primary and secondary distribution. The goal of such scheduling software will be to assure that production is handled efficiently without unreasonable delay, and that production errors, reruns, and bottlenecks are correctly identified as they happen.

Since we anticipate receipt of nearly 470 Census source tapes (from STF 1-4), which will be in turn compressed and archived onto about 200 archive tapes (with 200 backup tapes), a significant portion of report production will depend upon effective tape library management. STF data logging, installation, compression, and archiving, as well as report scheduling can probably be handled with a scheduling system written as an application of the Datatrieve data storage and retrieval system.

During Fiscal Year 1981 a decision was made to utilize the Datatrieve system for handling production scheduling and record keeping for the 1980 Census project, and a requirements analysis was made for this record keeping database.

7. Enhancement of SEEDIS System Software for Census Storage

The ability of LBL to undertake the effort of 1980 Census processing, data installation, and retrieval and report production is predicated upon the continued availability and maintenance of SEEDIS system software which has been developed over the past few years. In addition, certain enhancements are essential to SEEDIS in order for the Census Project to continue and achieve results within budget. These include:

- extension of the data model to handle multi-dimensional matrices and missing data elements in a convenient and comprehensive fashion;

- development of local network connections and archival capabilities on current LBL computer center mass storage hardware, as well as planning for future mass storage hardware advancements, should they come onto the marketplace

within the Census Project production cycle.

Extension of the Data Model: During the past year, significant progress was made in defining those aspects of the 1980 Census metadata which could not fit into the current SEEDIS data model, and the ways in which the data model could be enhanced to accommodate the new data files. A design document was prepared covering a range of future enhancements which would cover not only existing and anticipated files from the 1980 Census, but also future data files which are released in hierarchical data format, such as the Current Population Surveys.

Local Computer Network Connection: During the Fiscal Year, an outline of potential connections together with a quantitative examination of the capability of each type of connection to fulfill the archival retrieval requirements of the 1980 Census from the existing Automatic Tape Library mass storage device was prepared. One of the connections, the DGATE, was installed and has been operational since May 1981. It has been utilized to retrieve 1980 Census data below the census tract (census Block Group) level in metropolitan areas and below the Minor Civil Division level (census Enumeration District) in rural areas.

8. Acquisition and installation of Census Data and Map Files

LBL will acquire, install and test Census Summary Tape Files as they become available. Installation software will be tested originally with the preliminary tapes from the trial Census in 1979 in Richmond Virginia (the Richmond Test Tapes).

Geographic base files (map outlines) will be acquired and installed as they become available. They will be tested with existing mapping software and compared visually with equivalent or similar 1970 map files.

Machine readable data dictionary files from the Census Bureau will be acquired with the data tapes and a procedure to translate from them to SEEDIS data dictionaries will be applied as part of data installation.

During Fiscal Year 1981, the preliminary Richmond Test files from Summary Tape File 1-A,1-B were installed in SEEDIS and tested with a variety of installation software and report generation software. Data from the P.L.94-171 (Reapportionment population by race and ethnic origin) have been installed in SEEDIS to the Census tract level of geography, and archived on tape to the enumeration district/block group level. We have converted and installed the CENSPAK retrieval software which will read Census Data Dictionaries and may be used for testing data files, as well as understanding the data model concepts and transformation of Census Data Dictionaries to SEEDIS Data Definition Files.

A preliminary examination of boundary changes for areas tabulated for the 1980 Census shows that county boundaries have changed significantly for the states of Alaska and Virginia, and that there is a single county merger in the state of South Dakota. At the level of Census Tracts, the problem is substantial; nearly 30 percent of census tract boundaries have changed since 1970, either through splitting of tracts into a multiplicity of new tracts, or through the addition of SMSA counties and the creation of new census tracts. LBL has developed a methodology for aggregation and disaggregation as part of SEEDIS, to achieve comparability from 1970 to 1980 data. This methodology can be used to transform 1980 data to 1970 boundaries for mapping using the

existing boundaries.

9. Report Production

The process of report production consists of retrieval of data from SEEDIS installed databases for specified geography, aggregation to other levels of geography as desired, computation of derived values, and insertion of values into report templates which are written to tape for driving the typesetting printer.

During Fiscal Year 1981, LBL conferred in detail with the National Technical Information Service about the nature and components of report production and distribution, and the attached costs of such a production operation.

The Labor Market Projections Model

The aim of the Labor Market Projections Model (LMPM) is to provide estimates of current requirements for employment and training services in local areas. The LMPM project was developed in collaboration with the Employment and Training Administration of the Department of Labor to provide a comprehensive modeling system to be used by local CETA analysts. Since the LMPM project started at LBL in the summer of 1975, two models for estimating "current" local needs have been developed. Each year, both models are run for all states, SMSAs, prime sponsors, and other county groupings as requested by certain states. The results are mailed to the State Employment Security Agencies (SESAs) who then distribute them to the prime sponsors within each state. The information is utilized in the Annual Planning Information reports produced by each prime sponsor.

As well as developing and maintaining these two models, LBL has also looked at new data sources - for possible inclusion in one or both of the current models, or as a new means of providing relevant data for the SESA analysts. These new data sources include the 1980 Census, the Employment Security Automated Reporting System (ESARS), the Continuous Wage and Benefit History (CWBH) data base, and new uses of the Current Population Survey (CPS).

CPS-Census Updates

The first method, which utilizes Current Population Survey (CPS) data, provides economic indicators for the population in general as well as demographic population breakdowns. Updated Census tabulations are computed by multiplying the 1970 Census data by the ratio of the latest CPS data to the 1970 CPS data. Essential to this computation is a careful comparison of Census data categories with CPS data items. Tabulations equivalent to the Census tables are then performed on the CPS data. The difficulty with this procedure is that regional data is being used to update local area data.

In the past, the CPS updates were run for 21 regions across the United States - 11 states and 10 combinations of states. This was the greatest detail that was available when using the 1970 and 1975 CPS samples together. However, the CPS sample size was recently enlarged so that is is now possible to obtain state level data for every state. Now the CPS updates are run for 30 regions - 19 states and 11 combinations of states. This is the greatest detail available in the 1970 CPS. After the 1980 Census, it will be possible to run the CPS updates for each of the 50 states.

Population, Labor Force and Unemployment Projections

The second method uses a refined cohort-component method to project population, labor force and unemployment for each geoarea of interest. This involves the separate projection of births, deaths and net migration to give population by race by sex by five age cohorts for a particular area. Submodels are used to handle the military and collegiate populations separately. Using total population estimates for the target year as received from the Bureau of the Census, these cohort projections are then normalized to add up to a control total for that area. The labor force and unemployment projections are based on the labor force and unemployment of the geoarea in the base year, national changes in the labor force and unemployment from the base year to the target year, and the population projections for the target year as calculated above.

During the past year, work was done to make the results more accessible and useful to the SESA analysts. The model was converted from the Control Data 6600 in the LBL Computing Center to the VAX-11/780. As a first step, the data base upon which the projections model depends was installed in SEEDIS. By using the various graphing and mapping routines in SEEDIS, some or all of the data can be easily displayed according to user's specifications. Furthermore, this data can also be combined with any of the other data available in SEEDIS - the 1970 Census or LAUS, for example. The programs that do the projections calculations are also now operating on the VAX. Once a user friendly interface has been written, they too will be installed in SEEDIS.

Substantial improvements were made to the population projections model - enhancements were made to the underlying methodology, new data from the 1980 Census was installed in the model, and the format of the resultant print-outs was improved.

(1) A migration submodel was developed for geoareas that are counties or sums of counties. It estimates current migration flows by using total net migration estimates and the pattern of the area as determined by the 1960-1970 net migration rates. The document, "A Computerized Model to Estimate Current Net Migration Rates by Race, Sex, and Age", was written to describe the methodology used in this submodel. It will be mailed out to the state analysts this fall along with the requests for inputs for projecting to 1983.

In August, 1980 Census data on the population by age at the state level was received. These numbers were compared with the 1980 population projections as calculated by the LMPM projections model. The model was run twice - once using the new migration submodel as mentioned above, and once with the commonly used plus-minus adjustment technique.

The accompanying figure shows the results of this comparison. The first two columns indicate the mean absolute percent error over all age groups. The two remaining columns present the percent error for the age group having the greatest relative difference between the projection and the census result. The mean absolute percent error of the migration model is less than that of the plus-minus technique in all but two states. Furthermore, it is usually less by a factor of two. Looking at the maximum error, again the migration model outperforms the plus-minus technique and usually by a factor of two. For several states, Illinois, Louisiana, Montana, New Hampshire, New Jersey, and Texas, the percent error of the worst fitted age group using the migration model is less than the mean absolute percent error

using the plus-minus technique.

Further testing of the migration submodel will be done as more Census data becomes available.

(2) Several routines were written to force the sum of the population over each race/sex/age cohort in substate areas to add up to the corresponding race/sex/age cohort at the state level. They work well and converge rapidly to a solution even when forcing the race/sex/age cohorts in over 25 substate areas to add up to the corresponding figure for the race/sex/age cohort of the state. Once the projections have been successfully calculated for the largest substate pieces, the program subdivides these pieces where desired using the just calculated race/sex/age projections for each piece as control totals.

(3) Work was done to investigate the desirability of adding a third time interval to the projections. For the past few years, the projections model has projected in two time intervals - first from 1970 to 1975, and then from 1975 to 1980. The desired projections for 1981 or 1982 were then obtained by applying the 1981 or 1982 control total to the projected 1980 population distribution. It was felt that this introduced fewer errors than trying to project to 1985 and then interpolating between 1980 and 1985 to obtain 1981 or 1982. To be able to project to 1985, it is necessary to estimate fertility, mortality and migration rates for each area by race, sex and age from 1980 to 1985. Trying to estimate net migration by race, sex and age is very difficult.

However, this year the projections model will project to 1983. It was felt that it is not acceptable to apply a 1983 population control total to a 1980 population distribution. This distribution can change considerably over three years - particularly in areas that are growing rapidly or declining in population. Thus a third step was added to the population projections model. To project from 1980 to 1985, national mortality rates are used. State level fertility rates are obtained by a weighted average of the 1975-1980 state fertility rates with the national 1980-1985 fertility rates. The procedure used to estimate the 1980-1985 net migration rates changed considerably once the 1980 Census figures on population by race were added to the model. For more details on this please refer to subparagraph 5).

(4) The preliminary 1980 Census Population counts were aggregated as necessary and installed in the projections model. This aggregation was accomplished with a slight modification of the geocode file and a corresponding modification in one of the Codata tools. These preliminary figures, population by race for 1980, are used as control totals for 1980.

Since these figures were felt to be of general interest, they were also installed in SEEDIS as a separate data base. Previously they had been installed at the state, county, city, place, and MCD level. In this new data base, they are now also available at the prime sponsor level.

(5) As was discussed above, projecting to 1983 necessitated the addition of a third time step. The desired figures for 1983 are obtained by projecting first to 1975, secondly to 1980, thirdly to 1985, and then interpolating between 1980 and 1985 to obtain 1983. Since only a total population estimate for 1985 is available, it was thought that the migration rates for 1980-1985 would have to be estimated using the plus-minus technique. The migration submodel can be used only if race/sex population control totals for 1985 can be obtained or successfully estimated. However early testing of this idea shows severe problems. This is due mainly to the considerable change in racial patterns in some areas between 1970 and 1980. For example, in Prince Georges County, Maryland, the white population declined 30.4% from 1970 to 1980, while the black population increased 171.4% over the same time period. The plus-minus technique essentially extrapolates this trend to 1985 with disastrous results. It appears that it would be better to try to estimate the

1985 population by race and sex and then use the migration model to estimate 1980-1985 migration.

(6) To make the calculation of the unemployment projections consistent with the labor force projections, a CPS adjustment factor was introduced. The use of this factor adjusts the March-April 1970 unemployment rate of the area under study to the 1970 "annual" CPS. The CPS is a widely used and more accurate source of labor force data than is the Census.

(7) Checks were installed to the unemployment projections to insure that the unemployment rates were always below 100%. It was possible for the projected unemployment rate to be above 100% in areas with very sparse data and/or unusually high unemployment for a particular race/sex/age cohort in 1970.

(8) The projections model has always relied upon state provided inputs to calculate the labor force and unemployment projections. To increase the flexibility of the model and to take advantage of the knowledge of the local area analysts, several optional inputs were added this past year. Furthermore, the analysts may now submit up to three labor force and unemployment control totals. This allows the analyst to compare "high" unemployment with "medium" and "low". Many state analysts took advantage of this increased flexibility and submitted the optional inputs.

(9) Many pages of printed output are produced for each area for which the model is run. The contents of this output were redesigned. As in the past, tables are printed showing the input data, the intermediate calculations, and the resultant projections. This portrayal allows the analyst to trace what led to a particular result if so desired. Two pages of summary tables were added to the package. With these summary tables the analyst can easily compare results for various race, sex or age groups. Furthermore, the age cohorts have been redefined to agree with those specified in the CETA Forms Preparation Handbook.

With the changes outlined above, the population projections model has been greatly strengthened. As a result, the labor force and unemployment projections are similarly improved.

The computer programs for both the CPS-Census update model and the Projections model exist as batch programs. LBL "runs" the models each year and mails the results to the SESAs. Furthermore, during the past year, LBL has processed special requests for approximately 20 states. Some of the requests were for new county groups within the state; the remaining requests were to use new control totals for previously defined areas.

CPS Feasibility Study

LBL collaborated with Malcolm Cohen from the University of Michigan in preparing a publication containing estimates of the 1980 labor force and 1980 CETA population. These estimates were based on the March 1980 Current Population Survey (CPS). The collaborative effort took advantage of the microdata processing capabilities at the University of Michigan and the graphing and mapping capabilities of SEEDIS at LBL.

Once the tapes containing the CPS were received by the University of Michigan, they were tabulated to provide estimates of both the labor force and the CETA population by various demographic characteristics. Using the network access machine (NAM) on the CSAM UNIX machine, these estimates were transferred from Michigan to the DCN machines. Once on the DCN, the graphing and mapping routines from SEEDIS were used to display these estimates. The resultant maps and bar charts were mailed back to Michigan for inclusion in the publications that were mailed out to each of the fifty states.

ESARS Data Feasibility Study

Some time was spent assessing the merits of including portions of the ESARS data in one of the existing models in LMPM. Both models in LMPM are estimating current demographic conditions based on "older" local data and more current survey or summary data. ESARS data are relatively current and exist at the local level. However, the models presently in LMPM and ESARS are measuring two different things. In LMPM, the models are estimating characteristics of a population at one point in time, eg., how many white females aged 20-24 in the labor force are unemployed as of July 1, 1982? On the other hand, by tabulating ESARS data, one can get an estimate of how many white females aged 20-24 were unemployed during 1982. The estimates that can be obtained by using ESARS data are not directly compatible with those provided by LMPM.

Such estimates as could be obtained from ESARS may be of great interest to local planners and could complement the estimates already in LMPM. LBL already has considerable expertise with ESARS data as it receives state level summary tabulations from all the states. The information from these tables is available in an on-line retrieval system. In order to provide local estimates of the population in need of manpower services over a time period, the states would have to include a new substate level tabulation in their regular runs.

Investigation of CWBH Data Base

The CWBH database is a collection of employment and unemployment related data from sample surveys of the insured employed and unemployed in each state. It was felt that this was a potentially valuable source of current unemployment data at the local level. In order to get a better understanding of exactly what was in this database, a tape containing data for the state of Pennsylvania was obtained. The sampling rate in Pennsylvania is 1%. The data on this tape was for the time period April 1979 to September 1980. Several frequencies and cross tabulations of demographic items were obtained using SPSS. After a brief look at this data, it was judged that the data is of limited use to the LMPM for the following reasons:

(1) The ethnicity code excludes Hispanics in the White and Black categories while the LMPM include Hispanics in one of the racial categories - White, Black, or Other. Unless, Hispanics constitute only a small proportion of the population in the state, the race in LMPM can not be compared with ethnicity in CWBH.

(2) CWBH covers only the unemployed who are eligible and file for unemployment claims while LMPM includes all unemployed regardless of the eligibility for UI benefit. Thus unemployment rates using CWBH data are underestimates of the true rates. Furthermore, the eligibility requirements for UI benefits vary according to the individual state laws. It is thus necessary for the users to be familiar with the UI programs in each state in order to avoid possible misinterpretation of the results.

(3) Currently, CWBH data are only available in 14 pilot states. Since the scope of LMPM is the entire United States, the usefulness of the CWBH data in LMPM is limited at present.

(4) The CWBH data is a microdata file that includes a sample of UI claimants who filed for UI benefits. The number of UI unemployed thus covers a time period (eg., July 1979 - June 1980) rather than the point estimate provided by the LMPM (eg., July 1, 1981). The two estimates of the unemployed are not directly comparable.

Despite these problems, the CWBH can potentially be used as validity checks of the LMPM results in terms of order of magnitude and consistency of the projected numbers. The CWBH data does provide UI unemployed data down to the city and county level. However, a fair amount of administrative and data processing effort would be needed to obtain and analyze the CWBH data and then perform the comparison with the LMPM results.

ESARS Data Retrieval Program

The purpose of this project is to provide mechanisms for the storage, retrieval, and display of ESARS data at the state and regional level. This project is being undertaken on an interim basis until the redesign of the ESARS data is implemented by ETA.

At the start of FY81, the DCN-ESARS system completely replaced the BKY-ESARS system by performing all the basic functions of data reduction, data retrieval, and report generation necessary for production of ESARS reports.

The major accomplishments for FY81 were the release of ESARS Version 2 to Regions IX, X, and DC, the release of ESARS Version 3 to Regions IX, X, and DC.

With the installation of these versions on the regional nodes, the following major features became operational:

(1) Expansion of database coverage to include the following tapes:

BKY-standard Nation-by-State tape,
National-Regional Totals tape,
D.C.-standard Nation-by-State tape.

Handling of this last tape will obviate the necessity to convert historical ESARS data from the BKY system.

(2) Additions to database-handling modules to guarantee the correctness and stability of the ESARS database. Such additions include the acceptance of revised tape submissions from the states and consistency checks in storage and retrieval modules.

(3) Creation of a set of database-management modules for the ESARS Managers. The set provides database inventory capabilities, as well as control over database contents.

(4) Conversions of ESARS database tapes to a new format to simplify extractions from tape and to enable smoother recovery procedures for corrupted ESARS database tapes.

(5) Implementation of a shareable FORMATS library within the ESARS database, including automatic format file retrievals from the library by the ESARS report writer, and new user commands to simplify use of the library. Also, a new ESARS command was created to aid in the design of format files.

(6) Creation of a module to provide user-controlled selective checks on availability of ESARS data within the database, by geographic area and time period. The same module also checks on availability of format files in the FORMATS library.

(7) Enhancements to the ESARS report writer, including access to quarterly data and the usage of "NOT AVAILABLE" codes to allow processing of reports past non-retrievable items.

At the start of the fiscal year, work had already begun on ESARS experimental version 2 as follows:

The "report generation" (REPORTGEN) module had been changed to recognize and satisfy requests for quarterly data. It also provided a more forgiving error recovery for such user errors as divides by zero, and more precise diagnostics for incorrect or incomplete "format" files. Its dataset retrieval process had been enhanced to check tape lists as well as disk lists so that the user could know the availability and location of the datasets needed. The module could recognize

and replace old-style BKY CHART numerical codes such as 4E50 in "format" files and replace them in the resulting "script" files with the appropriate VAX CHART numerical codes to avoid numerical overflows in the VAX.

The "data reduction" (NEWDATA) module had been extended to handle the "Nation States" tape and the "National/Regional Totals" tape in their current formats. It had also been restructured to reduce the time and effort required to process several individual state tapes (MA462R1) coming in at one time into the DCN-ESARS database. All database modules had been modified to be able to accept revised tapes from the states containing amended versions of data sets.

During the first and second quarters of FY81, work was completed on a group of utility modules that complement the "data reduction" (NEWDATA) module and allow the Esars Manager of any regional DCN-ESARS node to directly control which portion of his local ESARS database resides on disk at any one time. The new commands and modules were extensively tested and the on-line documentation and assistance for the ESARS Manager was updated to include information pertinent to the new modules and commands. The command names are as follows:

EXTRACT
Will extract to disk datasets (and/or subsets) of the Manager's choice from the DCN-ESARS local database tapes

REDUCE
Will reduce a dataset and/or subset residing on disk to a smaller subset of the Manager's choice

REMOVE
Will remove (delete) from disk datasets and/or subsets of the Manager's choice

In order to insure the consistency of the DCN-ESARS database, all tape-handling modules were amended to recheck writes to DCN-ESARS database tapes, and on a read from such a tape, to read in the last amended version, if any, of a dataset.

At the Esars Managers' request, a new switch was included in the "data reduction" (NEWDATA) module to allow the Manager to suppress echo prints of tables found in the incoming dataset.

Work was completed on the CHECK module which enables the DCN-ESARS user to check the availability of datasets and/or subsets in the local DCN-ESARS tape and disk database. All checks are user-selected, by geographic area and time period.

Pursuant to the SEEDIS-CHART designation of a CHART standard, ESARS was made to call the CHART standard.

Two help commands and their on-line descriptions were added to the general user's menu. These are:

DESIGN
to help the user design his own format files,

VERSION
to describe the new features in the latest ESARS version.

The VERSION command was also added to the ESARS Manager's menu.

A 'news file' was added to the ESARS login procedure for programmer-to-user communication. A similar 'news file' was added to the ESARS Manager's login procedure.

The DCN-ESARS Version number, date of run, and timing prints were added to all modules.

An effort was made to standardize and comment all DCL command files and fortran codes so that they are self-documenting. At a later date, more

comments may be added for further documentation.

The release of DCN-ESARS Version 2 to any one regional VAX node became dependent on the previous installation of VAX/VMS Version 2 in that node. The testing of all new modules and features was completed, and it was verified that the old modules and those currently under development still worked correctly under the new VAX/VMS operating system.

The format for ESARS database tapes (also called ESARS Master tapes) was changed to allow more efficient extractions from tape and better procedures for recovery of corrupted ESARS Master tapes. The scheme for accepting revised tape submissions was updated to conform to the new ESARS Master tape format and modules that add to the ESARS database on tape and those that read from it were updated to conform likewise to the new ESARS Master tape format. Just prior to installation of ESARS Version 2, all existing ESARS database tapes (for Region 9: ST79,ST80,ST81; for Region 10: ST80,ST81) and their backups were converted to the new format for ESARS Master tapes.

With the installation of ESARS Version 2, all the features described above became operational.

Work on ESARS Version 3 progressed during the third and fourth quarters of FY81, culminating in its release just prior to the end of the fiscal year. With the installation of ESARS Version 3 on the regional nodes, the following additional features became operational:

The commands and modules needed to implement the shareable library of report formats and descriptions were created, tested, and documented. All general users will be able to add their format files into the library of shareable formats. The library forms part of the local DCN-ESARS database on disk, and resides in directory SY$ESARSDAT:[ESARSDAT.FORMATS]. The commands that handle the FORMATS library are shown below:

Command	Parameters	Optional Parameters
ADD FORMAT	filename	(newname)
ADD FORMAT(s)	filename(s)	
CHECK FORMATS		(filename(s))
CHECK FORMAT NAMES		(filename(s))

A sample format file was chosen for inclusion in each shareable FORMATS library and informational messages regarding its existence were added to the ESARS news files and to the on-line HELP. The REPORTGEN module (REPORT and TRANSLATE commands) was changed to automatically look for a requested format file in the shareable FORMATS library whenever the file is not found in the current user directory.

Off-line printed documentation, a "Beginner's Guide to ESARS" was created and refined.

The ESARS user interface was changed

to accept abbreviations of all ESARS commands down to their first three letters, and parameters down to their first letter.

REPORTGEN now allows completion of processing of a format file past retrieval errors due to inaccessibility of data. All expressions containing non-retrieved items are flagged with a value of 2.E20 (interpreted by CHART as NOT AVAILABLE) and an error message identifying each item is printed on the terminal during processing. The REPORTGEN module was further refined to compile error messages due to non-retrievable data items in the order they appear in ESARS datasets to aid the user in finding and fixing errors in his format file. Certain repetitious informational messages identifying the data sets retrieved for report processing were deleted from the user interface.

The CHECK module was expanded to allow the user to list the tables contained by subsets in the ESARS database.

Database inventory modules were created, then fully tested and documented. They are accessible via the expanded CHECK command and may be used to list contents of the database according to areatype or year of interest or to print out the total size of the database. The new commands are shown below:

Command	Optional Parameters
CHECK ALL	(DATASETS)
CHECK STATEWIDE	(DATASETS)
CHECK NRTOTALS	(DATASETS)
CHECK 19yr	(DATASETS)
CHECK TOTAL SIZE	

Access to the following DCL commands and Software Tools was added to the ESARS Managers' modules:

$DIRECTORY
$TYPE
$SHOW
$HELP
$WHO
$TALK

Modifications were made to the NEWDATA data-loading module to clarify the selection of datasets/subsets to be put on disk, and print-outs of available disk space were added to the NEWDATA module. Improved error messages identifying dataset, table, line number, and cell where errors were found during reading and interpretation of incoming ESARS datasets were incorporated into the dataset print-outs provided by NEWDATA, to form a permanent record of errors encountered in incoming data tapes.

Off-line supporting documentation was installed, which describes the function of the ESARS Manager and the structure and philosophy of the ESARS database for which he is responsible. It also contains a detailed description of the interaction that takes place whenever he runs NEWDATA - the module that loads new data into the

DCN-ESARS database. The on-line HELP for the ESARS Manager was updated to include descriptions of the new database inventory modules callable via the expanded CHECK command. Descriptions of DCL commands and Software Tools accessible through the ESARS Manager's modules were also included.

With regard to the decision to convert historical data from the D.C. tape library rather than from the BKY library: enhancements made to the NEWDATA module to handle the new numeric formats and to recognize and screen out possible errors in the data will also enable the program to accept the non-coded ETA/SESA tapes produced by UNIVAC and Honeywell state computers. At the completion of successful testing on simulated input data as well as on data from a D.C.-standard Nation States tape, field testing of the module was begun in ETADC.

Illustrations:

Diagram of Seedis:

Figure 1 presents the logical structure of Seedis, with its underlying file interchange format, unified interface, and various functional modules. The schematic view indicates the major functional components of Seedis. The image was generated by the graphic editing capabilities of Seedis. In summary the capabilities of Seedis are 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 online documentation; and
efficient handling of large numeric databases.

Seedis Chart Examples:

The Chart examples shown on the following pages demonstrate analytical quality and presentation quality dot, line, bar, and circle (pie) charts that Seedis can generate. The images indicate the fineness of resolution for typography, lines, and gray values that can be available to users who are trying to determine the ideal format in which to present their data. The exact size, position, and appearance of titles, figures, and other elements are adjustable. The images come from several different hardware configurations in which Seedis can generate charts.

Figures 2, 3, 5, 6, 8 and **9** indicate the standard kinds of charts that Seedis can generate automatically. The images were originally developed on a Tektronix 4013 and displayed on a Tektronix 4631 hardcopy device. They were then redesigned for presentation as both color slide and black-and-white printed images. They appear here reduced in size to 50% of the original. Each image (including typography) was generated on a Dicomed D48 film recorder. They indicate the high quality of typography and line quality that is important when reproducing charts in final printed publications and in audio-visual presentations.

Seedis Mapping Examples:

The mapping examples shown on the following pages demonstrate the kinds of analytical quality and presentation quality area and symbol choropleth maps that Seedis can generate automatically. As with the preceding chart images, these examples have been specially redesigned with the interactive graphic editing capabilities of Seedis. These maps indicate the fineness of resolution of typography and lines that can be available to users who are trying to determine the ideal format in which to show their data.

Exact textures, symbol size and position, and typography are adjustable. Several different hardware configurations can be used to generate the images.

Figure 4 shows a bivariate area choropleth map of San Francisco for a geographic level of Census tracts. The bivariate choropleth map allows the viewer to determine locations of positive and negative correlation between two parameters. This kind of image can be 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 was originally developed on a Tektronix 4631 hardcopy device. The image was redesigned on a Ramtek color terminal for presentation as a black-and-white printed image of the size shown and generated on a Dicomed D48 film recorder as a 4" x 5" negative.

Figure 7 shows a symbol map that uses circle (pie) charts to show multiple population characteristics of California Standard Metropolitan Statistical Areas (SMSAs). Note the division of the circle symbol into varying gray values and the overlapping of circles with automatic erasure of underlying circle areas. Normally these images are drawn in color and are recorded on 35 mm. or 4 x 5 inch color film or on 8 x 10 inch Polaroid color prints. This example was originally developed on a Tektronix 4014 terminal and proved on a Tektronix 4631 hardcopy device. The image was redesigned on a Ramtek color terminal for presentation as a black-and-white printed image of the size shown and generated on a Dicomed D48 film recorder as a 4" x 5" negative.

Figure 10 shows a flow map produced by a special facility which, given a few intermediate points along the arrow path, generates smooth arrows between end points. This example was originally developed on a Tektronix 4014 terminal and proved on a Tektronix 4631 hardcopy device. The image was redesigned on a Ramtek color terminal for presentation as a black-and-white printed image of the size shown and generated on a Dicomed D48 film recorder as a 4" x 5" negative.

Figure 11 shows the results of the comparison of 1980 Census data on population by age with populations calculated by the LMPM projections model. The first two columns indicate the mean absolute percent error over all age groups. The two remaining columns present the percent error for the age group having the greatest relative difference between the projection and the census result.

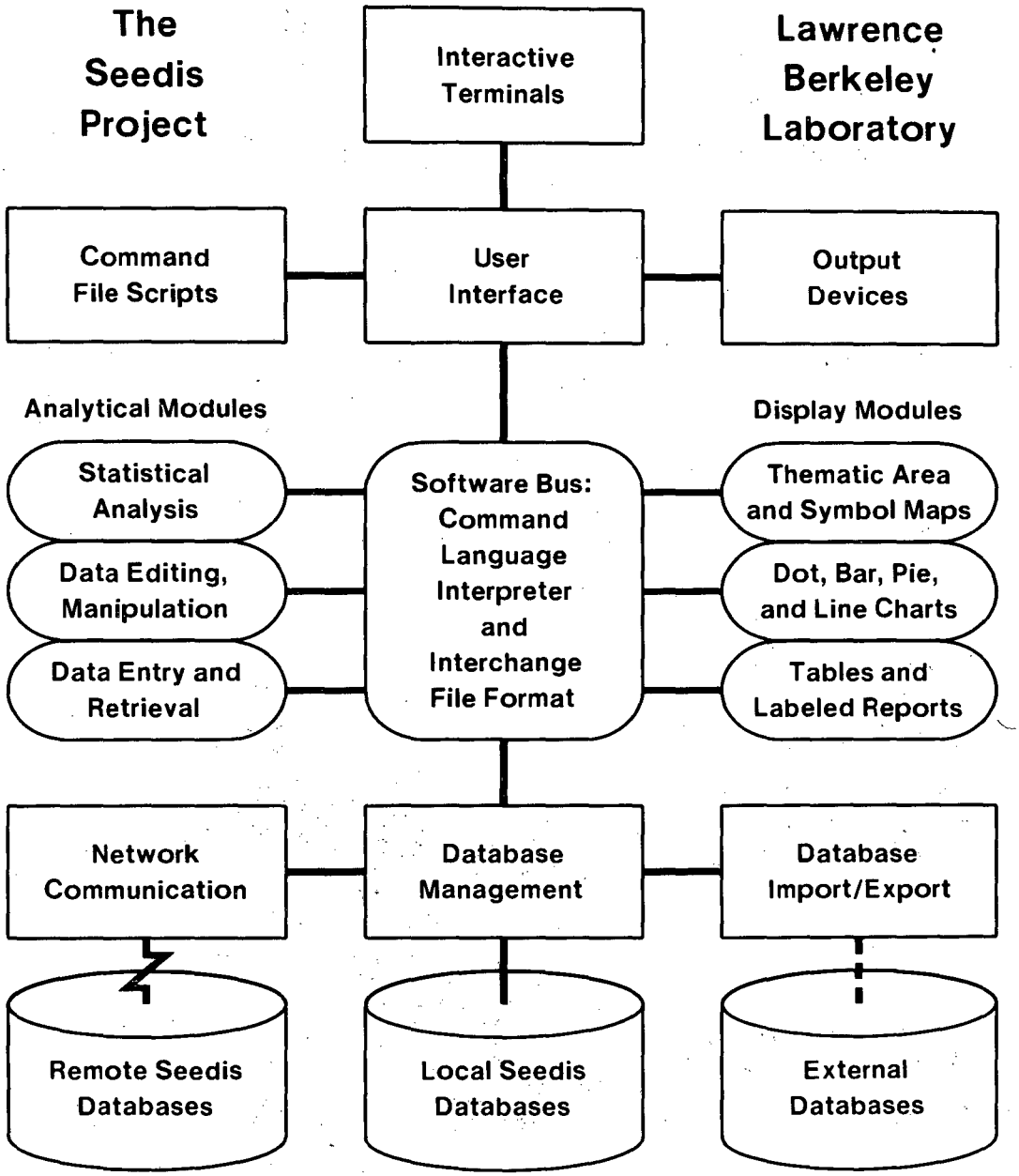


Figure 1

XBL 828-11046

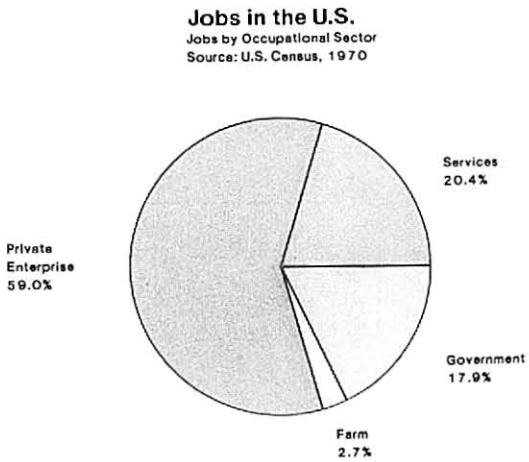


Fig. 2

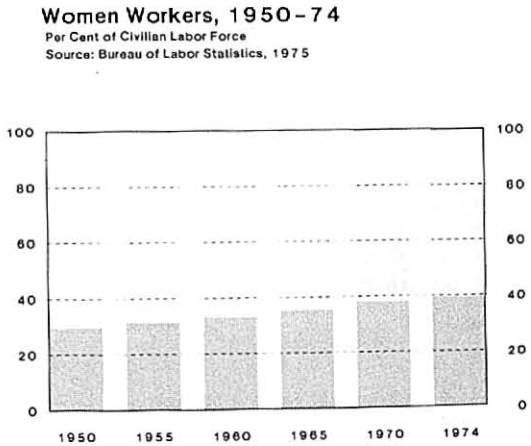


Fig. 3

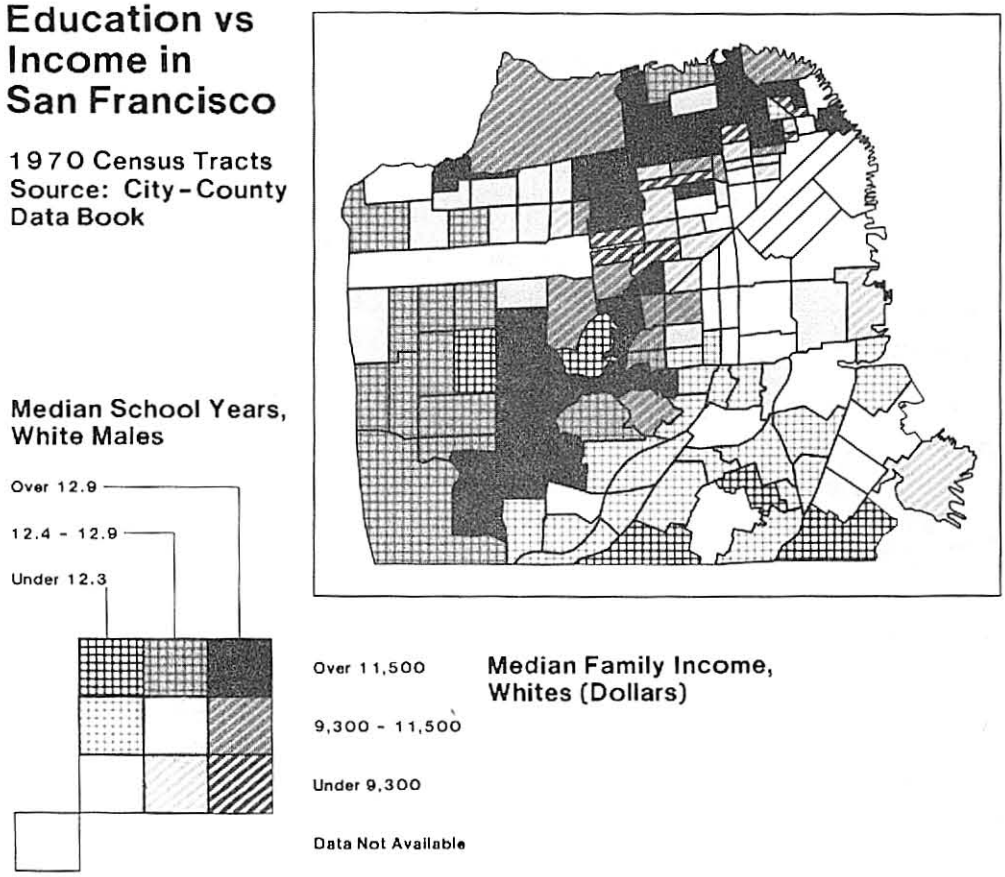


Fig. 4

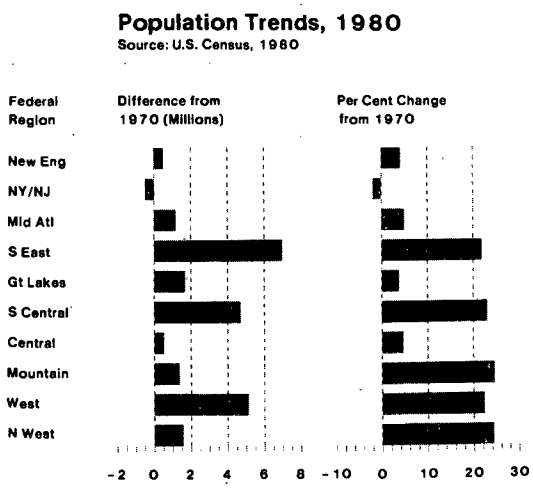

Population Trends, 1980
Source: U.S. Census, 1980
Federal Region
Difference from 1970 (Millions)
Per Cent Change from 1970
New Eng
NY/NJ
Mid Atl
S East
Gt Lakes
S Central
Central
Mountain
West
N West
-2 0 2 4 6 8
-10 0 10 20 30

Fig. 5

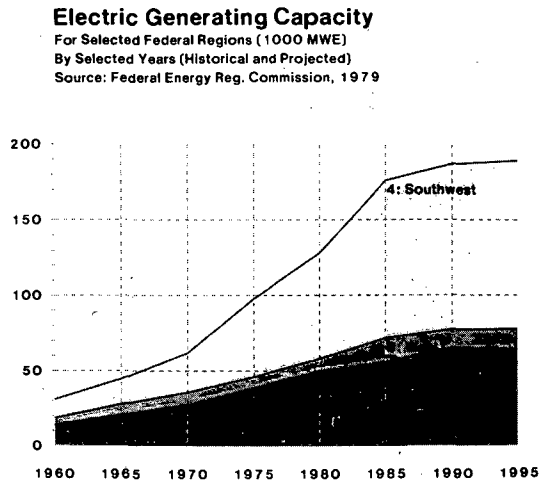
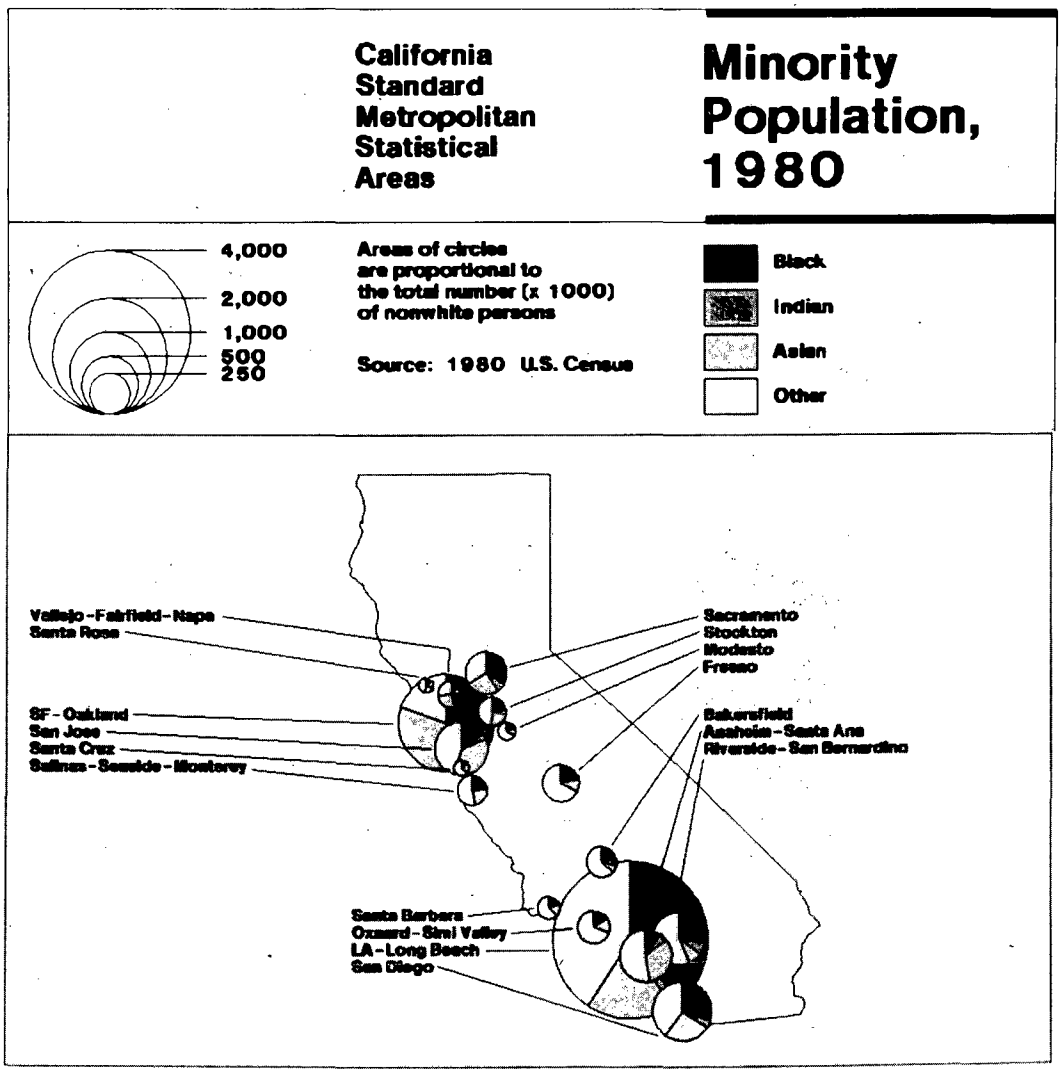

Electric Generating Capacity
For Selected Federal Regions (1000 MWE)
By Selected Years (Historical and Projected)
Source: Federal Energy Reg. Commission, 1979
4: Southwest
200
150
100
50
0
1960 1965 1970 1975 1980 1985 1990 1995

Fig. 6


California Standard Metropolitan Statistical Areas
Minority Population, 1980
4,000
2,000
1,000
500
250
Areas of circles are proportional to the total number (x 1000) of nonwhite persons
Source: 1980 U.S. Census
Black
Indian
Asian
Other
Vallejo-Fairfield-Napa
Santa Rosa
SF-Oakland
San Jose
Santa Cruz
Salinas-Seaside-Monterey
Sacramento
Stockton
Modesto
Fresno
Bakersfield
Anaheim-Santa Ana
Riverside-San Bernardino
Santa Barbara
Oxnard-Simi Valley
LA-Long Beach
San Diego

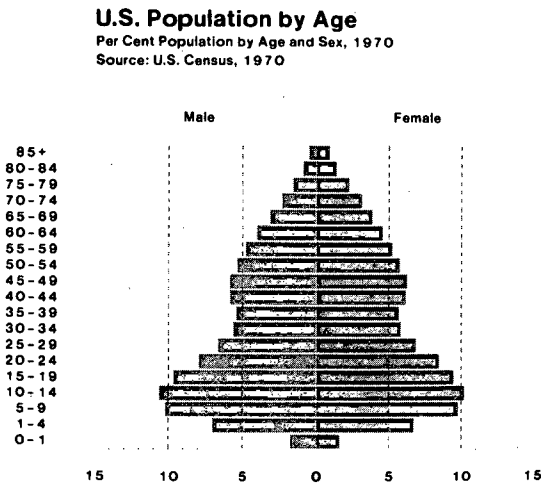


Fig. 8

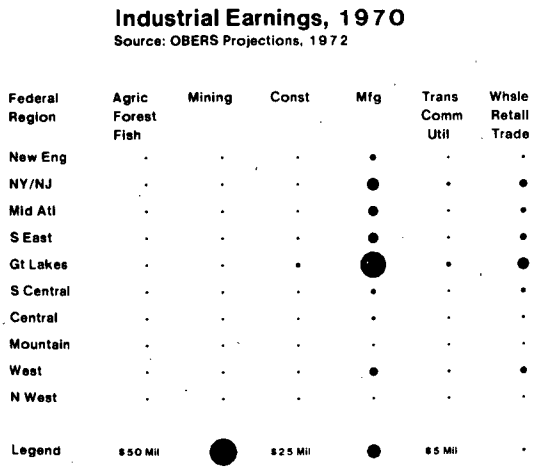


Fig. 9

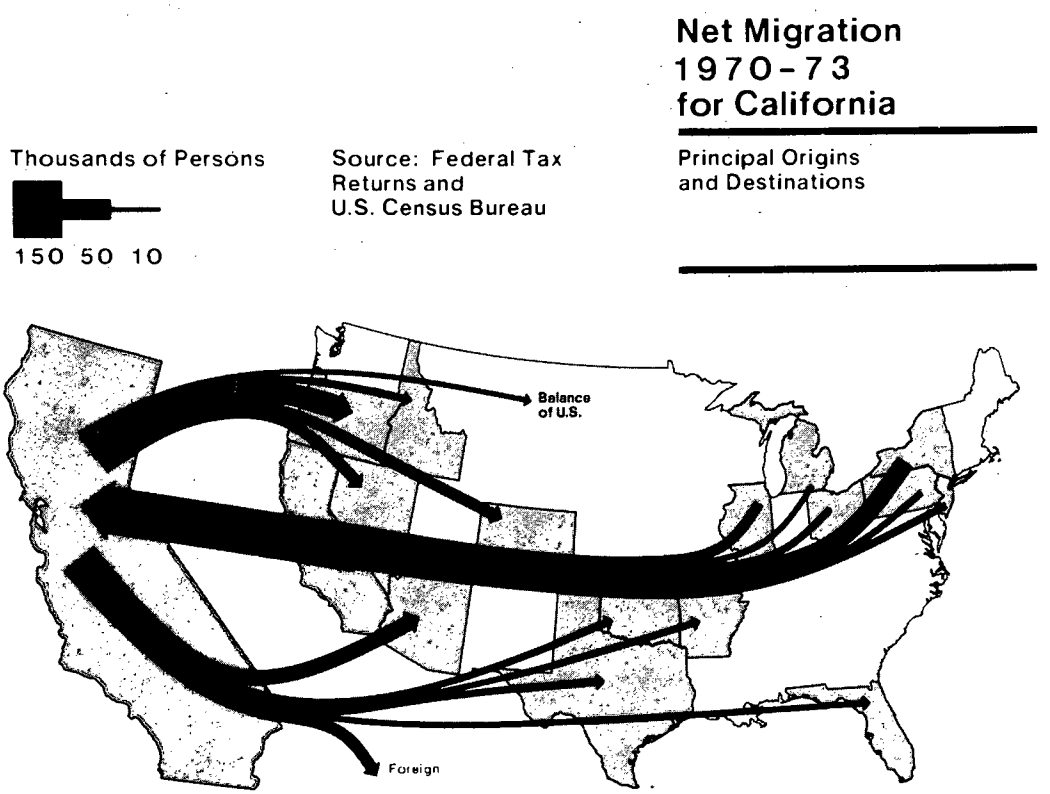


Fig. 10

Comparisons of 1980 Population Projections with 1980 state level Census data

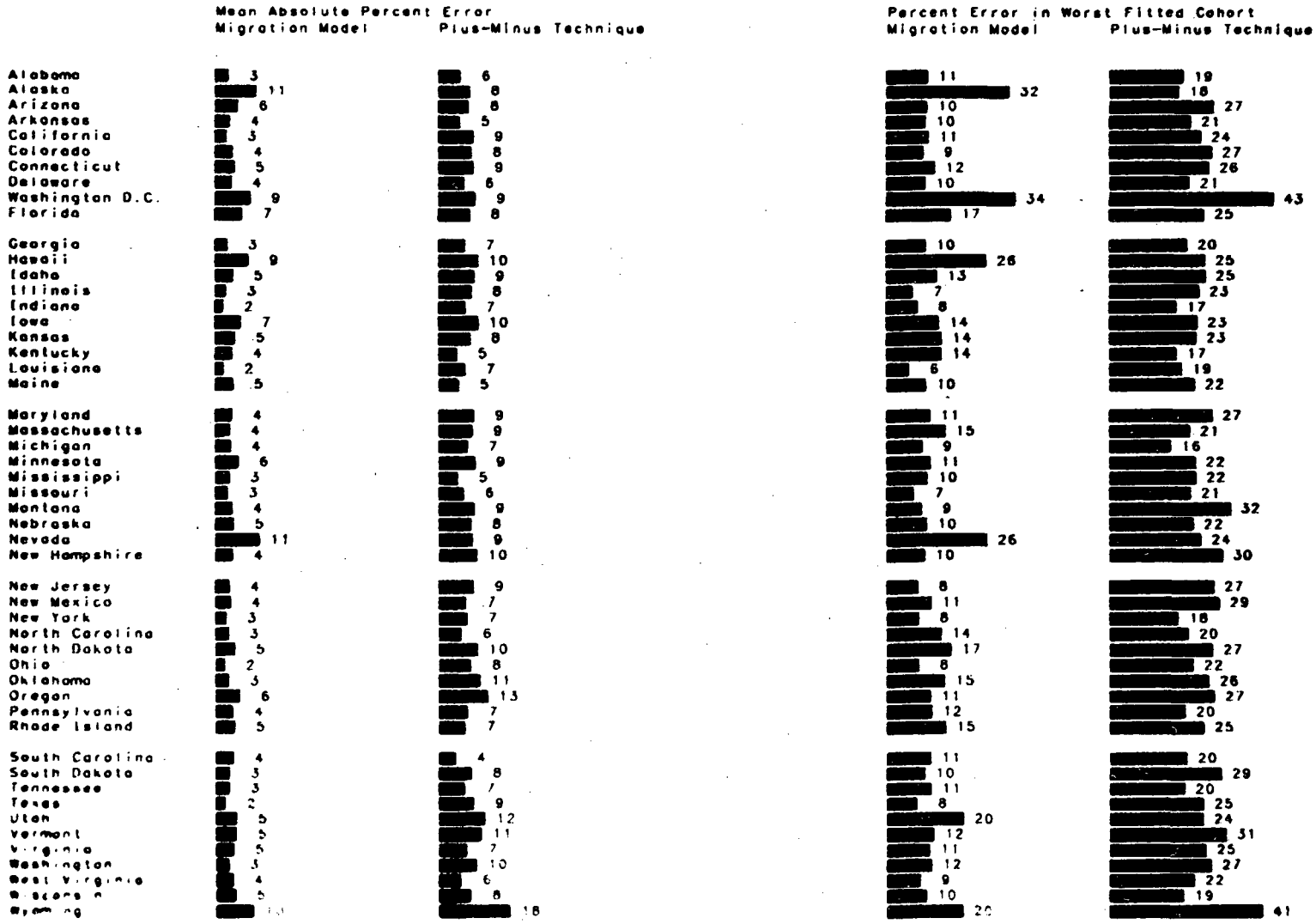


Figure 11

XBL 828-11047

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