AD-A 676 [illegible]

THE MAPPING OF DISEASE (MOD) PROJECT

THE GEOGRAPHIC DISTRIBUTION OF INFECTIOUS DISEASES

DA 49-092-ARO-130

FINAL REPORT

31 AUGUST 1968



COMPUTERIZED MAPPING OF DISEASE

USAREP / AFIP (Geographic Pathology Division)

Reproduced by the
CLEARINGHOUSE
for Federal Scientific & Technical
Information Springfield Va 22151

This [illegible] has been approved for [illegible] sale; its distribution is unlimited

THE MAPPING OF DISEASE (MOD) PROJECT

THE GEOGRAPHIC DISTRIBUTION OF INFECTIOUS DISEASES

DA 49-092-ARO-130

FINAL REPORT

31 AUGUST 1968

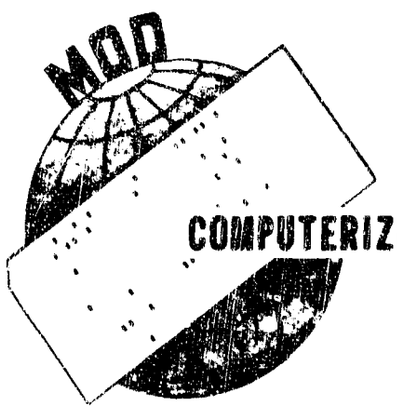
Prepared by: *Howard C. Hopps, M. D.**

*Roger J. Cuffey (Ph.D.), Capt., MSC, USA***

Jerome Morenoff, S.J.D.†

Wayne L. Richmond, A.B.†

*Joseph D. H. Sidley, M.S.**



COMPUTERIZED MAPPING OF DISEASE

UAREP / AFIP (Geographic Pathology Division)

* AFIP and UAREP

** AFIP

† PRC

Background information related to this contract

This information is presented under the following headings:

Contract data 0-1

Reports submitted 0-2

Objectives 0-3

Project plans 0-5

Advisory committee 0-7

Point at which project was terminated . 0-10

Fiscal data 0-12

A JOINT REPORT FROM

The Universities Associated for Research and Education in Pathology (UAREP)

The Armed Forces Institute of Pathology (AFIP)

This research was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Life Sciences Division, U.S.Army Research Office, under Contract N DA 49-092-ARO-130.

This "final report" is submitted in fulfillment of the Contract to Universities Associated for Research and Education on Pathology (UAREP).

Order Number	113.0761
Program Code No.	P6G20-25
Name of Contractor	UAREP
Date of Contract	15 Nov 1965 and 15 Nov 1966
Amount of Contract	83,734 and 84,434
Contract Number	DA 49-092-ARO-130
Contract Expiration Date	31 August 1968 (extended from 15 Nov 1967 without additional funds)
Project Scientist (UAREP)	Harlan I. Firminger, M.D. (Univ. of Maryland)
Associate Scientist (UAREP)	Howard C. Hopps, M.D. (AFIP)
Title of Work:	Geographic Distribution of Infectious Disease (Computerized Mapping of Disease -- MOD)

This report is also submitted in fulfillment of the following sub-contract to Planning Research Corporation (PRC).

Name of contractor	PRC
Date of contract	1 Feb 1967
Contract Number	6 1
Project Manager (PRC)	Jerome Morenoff, S.J.D.
Title of Work:	Computerized Mapping of Disease (MOD) Project Design, Implementation, and Developmental Testing

REPORTS SUBMITTED

The following reports were prepared in partial fulfillment of contract DA49-092-ARO-130 requirements during the course of the Geographic Distribution of Infectious Disease Project

(1) 28 Jan 1966, Contractor Quarterly (1st) Progress Report (15 Nov. 65 - 1 Feb. 66), 3 pp.

(2) 1 May 1966, Contractor (1st) Semi-Annual Progress Report (15 Nov. 65 - 1 May 66), 6 pp.

(3) 8 Aug. 1966, Contractor Quarterly (3rd) Progress Report (1 May 66 - 1 Aug. 66), 5 pp.

(4) 15 Dec. 1966, Contractor (1st) Annual Progress Report (15 Nov. 65 - 14 Nov. 66), 102 pp.

(5) 3 Mar 1967, Contractor Quarterly (5th) Progress Report (15 Nov. 66 - 15 Feb. 67), 6 pp.

(6) 15 May 1967, Contractor (2nd) Semi-Annual Progress Report (15 Nov. 66 - 15 May 67, 76 pp.

(7) 31 Aug. 1968, Contractor FINAL REPORT (15 Nov. 65 - 31 Aug. 1968), pp. 430.

In addition to the seven reports listed above (for external use) there were six other reports prepared for internal use: Four of these were prepared by Planning Research Corporation in fulfillment of subcontractor requirements. and two were prepared by members of our own staff: Capt Roger J. Cuffey (AFIP) and Mr. G. G. Gullett (U. of Ill.).

The final report, i.e., this volume, incorporates all significant (current) information that has been presented in all previous reports -- internal as well as external ones. (The internal report dealing with published maps that relate to ecology of disease, by Mr. G. G. Gullett, is reproduced in its entirety.)

OBJECTIVES

Present objectives of the MOD system, in the context of nearly three years experience with this program are presented in the Introduction (1.2). But there would be advantage in presenting our objectives at the time the MOD project began. These objectives were set forth in three paragraphs contained in our original application, dated 15 June 1965 They are reproduced here verbatum.

* * *

A. OBJECTIVE OF THE PROGRAM The ultimate objective of the program is to develop research methodology by means of which the occurrence of a particular disease may be correlated with a variety of sociological, physical and environmental factors such as population density, races, ethnic groups, altitude, temperature, humidity, character of the soil, agricultural products, possible insect vectors and animal reservoirs of disease.

The immediate objective of the program is to provide data in the form of disease distribution maps and atlases, showing prevalence, incidence, and severity of specific infectious diseases throughout the world.

B. TECHNICAL NEED FOR THIS PROGRAM Data on the geographic distribution of infectious disease are of obvious importance in evaluating the disease risk for groups of persons assigned to foreign posts and in any detailed planning that involves the socio-economic problems of a particular area. There have been only two major contributions in this field. They are: (a) Geographic Atlas of Disease, prepared by the American Geographical Society, published during 1950-55. (b) World-Atlas of Epidemic Diseases, edited by Professor Ernst Rodenwaldt (Heidelberg), published in 1952 but reflecting data gathered some years before. However, data on some developing countries of current interest are either sparse or completely lacking. The methodology developed by this program would provide a means for linking contributing and precipitating factors with a given disease thereby providing clues to the etiology of the disease and suggesting specific basic research for methods of control.

C. RELEVANCE TO ARPA MISSION Infectious diseases are the greatest cause of morbidity and mortality among troops and civilians in time of war. Infectious diseases also exert a major influence on the socio-economic status of all countries, especially developing countries. The proposed study would provide valuable information on the distribution, prevalence, and incidence of infectious disease throughout the world, and would provide a method for carrying out rapid and effective searches for important interrelationships among a large variety of potential causal factors of any given disease. The development of methods of control of infectious diseases is in accordance with ARPA/AGILE's mission of conducting research, development, and tests of techniques and equipment required by local forces in remote area confict situations.

PROJECT PLANS

To avoid redundancy, we direct the reader to the Preface, Section 1, Introduction (particularly the Definition of Goals, 1.21), Section 2, Technical Summary, and Section 9, General Summary, Conclusions and Recommendations, where project plans are considered in some detail.

In this statement, contributing background information about the project, we shall concentrate on relationships among biomedical professional members (principally AFIP staff) and computer scientist members (principally Planning Research Corporation Staff) of the MOD group.

At the outset it was realized that (as a generalization) those who understood disease ecology were not competent to direct a computer processing attack on the problem and that, conversely, those who were competent in the area of computer science/technology did not understand disease ecology. We concluded that, since computers were the means toward increased understanding (not the end), biomedical scientists should lead in development of the MOD system, seeking computer oriented specialists to support the very important computer processing aspects of the program.

Upon careful deliberation, and after it became clearly evident that our chances of finding a top notch systems analyst for short term hire on the open market were small indeed, we decided to employ the services of a commercial group to help us with this highly critical aspect of our program System Analysis. We met with representatives and considered informal proposals from four organizations:

Auerbach Corporation
Bunker-Ramo Corporation
* Planning Research Corporation
* Systems Research Group, Inc.

and received formal proposals from two of these*. In the course of our

deliberations we conferred with the AFIP Automatic Data Processing Section, the National Bureau of Standards, and the University of Illinois, getting advice and counsel in regard to the system which would be required for our computer-based automatic mapping program.

On 5 July 1966, Subcontract UAREP 66.1, awarded to Planning Research Corp. (PRC), became active. A critical factor in selecting this company was their agreement to assign two highly motivated and extraordinarily well qualified persons, full time, to our project. (Their very efficient work and the full cooperation of their associates at PRC more than justified our decision.)

The initial study uncovered many complex problems which were more serious than could have been anticipated before this study, and a second subcontract was let to PRC in August 1966 in order to complete analyses of these problems. Joint effort and close cooperation among the biomedical professionals of AFIP, and UAREP, and the data-processing professionals from PRC (and also one employed by UAREP) resulted in increased understanding of and tentative solutions to most of the problems which had been raised earlier.

Based upon the conclusion from the system analysis effort, namely, that a computerized disease-mapping system was technically feasible from the data-processing standpoint, the MOD project continued the services of PRC (by means of a third subcontract) and, beginning in February 1967, PRC and AFIP/UAREP scientists together conducted the system design effort for the proposed MOD system.

Implementation of that system, based upon the design accomplished, was to have begun in June 1967 and was estimated to require approximately another year. Following completion of implementation, a phase of system developmental testing, trial operation, and suitable modification of the MOD system would have occupied the remainder of the project's scheduled time. Instead, premature termination of ARPA's support required the MOD personnel -- both those from AFIP/UAREP and those from PRC to devote their remaining resources to completing system design and producing this final report.

INSTITUTE FOR DEFENSE ANALYSES

Science and Technology Division IDA 400 Army-Navy Drive, Arlington, Virginia 22202 • Telephone (703) 558-1000

Advisory Committee

This letter, photo-copied in its original form, explains the composition, functions and conclusions of the Advisory Committee.

7 September 1967

MEMORANDUM TO: Dr. Howard Hopps
Chief, Division of Geographic Pathology
Armed Forces Institute of Pathology
Washington, D. C. 20305

SUBJECT: Mapping of Disease Advisory Committee - Summary Report

I. Introduction

Following the recommendation of Dr. Herbert Pollack of the Institute for Defense Analyses and the concurrence of appropriate officials at Advanced Projects Research Agency and Armed Forces Institute of Pathology an advisory committee was formed for the Mapping of Disease (MOD) Program being conducted by the Geographic Pathology Division, AFIP. The membership of this committee is listed below:

Mr. Fred I. Edwards, ARPA
Mr. Ronald Finkler, IDA
Dr. Allan L. Forbes, ARO
Mr. Joseph E. Hinds, Consultant
Dr. Howard Hopps, AFIP
Dr. Myles Maxfield, Consultant
Dr. Donald R. Sheldon, IDA

The terms of reference for this committee were to clarify over-all governmental requirements relating to geographic aspects of disease and to assist the project director in assuring optimal responsiveness toward reaching these goals during the anticipated remaining 1.5 years of the original program plan.

Meetings of the Advisory Committee were held on 24 May and 13 June 1967 at the Institute for Defense Analyses. In addition, Mr. Finkler and Dr. Sheldon held separate meetings with Dr. Hopps and his staff on behalf of the committee. Their findings and recommendations have been incorporated in the body of this report.

At the outset it was envisioned that the Advisory Committee would play a continuing role in the development of the MOD program. However, as unexpected and irrevocable budgetary restraints have resulted in termination of funding for the program at the end of current obligations (November 1967), there seemed little need for continuing activities and this report is intended as final.

II. Findings

A. General Requirements:

1.0 The requirement for a greater understanding of the global and geographic aspects of disease extends throughout the official community and includes not only public health and military interests but also those associated with international development and foreign relations.

2.0 The scope of these interests range from basic considerations such as the etiology of infectious diseases to estimative or evaluative measurements of past or future events. Even such relatively gross measures as calculation of current incidence of infectious disease on a worldwide basis from crude data could benefit from appropriate application of computer technology.

3.0 A complete response to the total dimensions of the problem as defined above is clearly beyond the capability of monetary and personnel assets allocated to the MOD Program even if support had been continued at the projected rate. However, one application of computer technology which would be basic to all aspects of the problem would be a broad based data storage and retrieval system. In this regard, the experience and expertise developed by the MOD staff particularly in Data Extraction Format, Factors Cataloging and Data Structure Vocabulary would be a valuable input.

B. MOD Program Review:

1.0 In keeping with the initial funding proposal objective, the MOD Program concentrated on the design and assembly of a system for computerized mapping of disease information. The program plans and study methods appear reasonable and satisfactory for this purpose.

2.0 While the objective of providing an operational disease mapping system will not be met by the conclusion of the present contract (Nov. 67), much valuable information concerning how such a system might be developed has been derived.

3.0 The maximal lasting benefit obtainable from the MOD Program during the remaining period of support will be the preparation of a final report of their system analysis and design including a detailed evaluation of the problems associated with establishing the requisite data base for operation of such a system.

III Recommendation

The final report of the MOD Study should be directed toward providing a basic document which would serve primarily for the indoctrination and guidance of any future programs that might be directed toward the application of computer technology to epidemiological and geographic aspects of disease.

To meet this objective the following outline lists those aspects of the MOD Study which should be covered in the final report:

1. An introductory statement and discussion of objectives.

2. A summary of activities, project plans, and point at which project had to be terminated.

3. A summary of accomplishments, summary of contracts with outside organizations, and a list of all appropriate documentation and reference material.

4. The results and discussion of the experiments in contour mapping including the systems used, their limitations, examples of the outputs, and appropriate ways disease information may be presented on maps.

5. The complete system design specification including descriptions of all files and file structures, program flow charts and hardware configurations.

6. A complete description of:

   a) the data base(s) employed,
   b) the Data Extraction forms and instructions on their use,
   c) the Factors Catalog,
   d) the Data Structuring Vocabulary,
   e) the Query Language and other data,
   f) a glossary of appropriate terminology.

The assembly of the data and experiences of the MOD Project in a manner similar to that described would constitute a worthwhile contribution and should be accepted by the sponsoring agency as satisfactory fulfillment of contractual obligations.

For the Committee,[1]

Donald R. Sheldon

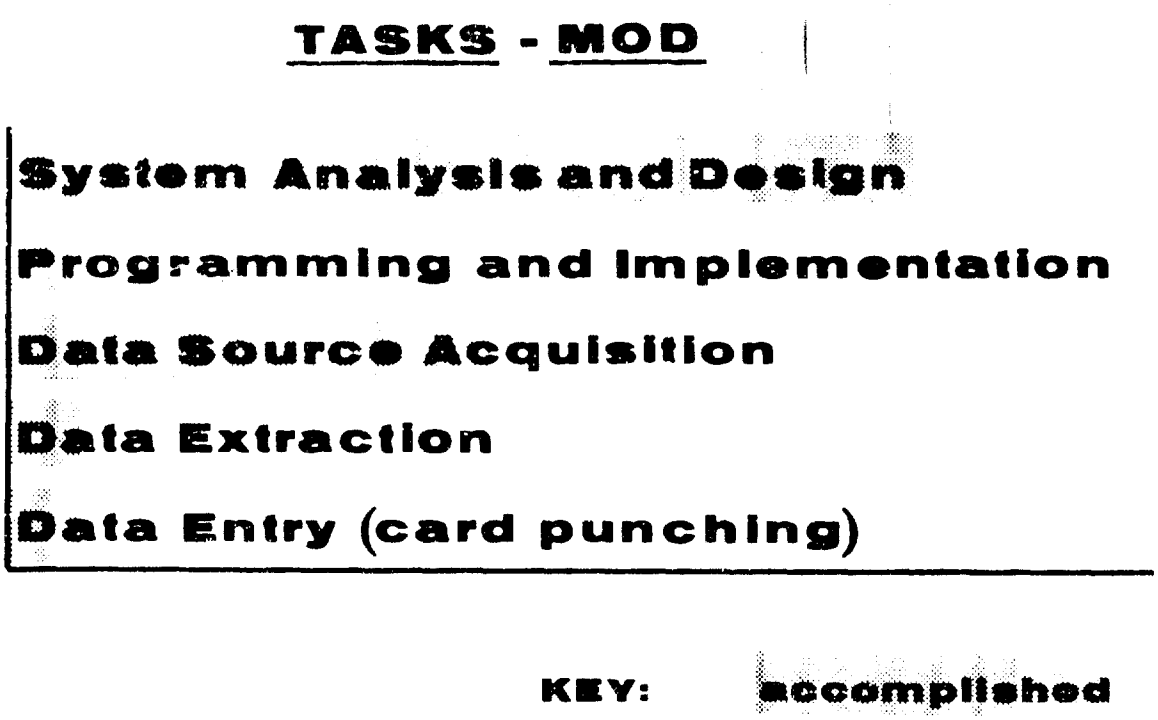
[1]Concurrence in Draft, 15 August 1967

POINT AT WHICH THE PROJECT WAS TERMINATED

At the outset, the MOD project was visualized as a three-year effort, and it was planned that the first two years would be primarily concerned with system analysis and system design. Implementation was planned for the third and final year. When it was learned that ARPA's support of the project was to terminate at the end of two years, all efforts were directed toward finishing data analysis and data structuring (method design), and completing the various aspects of the system design phase. (In the interests of efficiency, we had begun to implement some portions of the system so that we could more effectively design other portions.) It was our feeling, and one supported by the Advisory Committee, that a completed system design would represent an important milestone and would contribute information of value to other groups interested in comparable problems or, perhaps, in taking up where we left off in developing a system for the computerized mapping of disease-environmental data.

The system analysis and design have both been completed. (There are several aspects of the system design that will need further elaboration, but this cannot be performed outside the context of a partially implemented system.) In addition, we have made an extensive analysis of data characteristics: sources, limitation of the data, per se, and problems involved in preparing these data for computer input. A method for structuring data has been designed and tested, and a comprehensive factor catalogue has been produced. We have gained new insight into the characteristics of disease-environmental data that allow them to be mapped, and have developed data extraction forms reflecting these requirements.

In essence then, the MOD project was terminated just short of the implementation phase. The figure below shows the extent to which the five major component tasks were completed.



FISCAL DATA

Support for the biomedical portion of the MOD effort has been provided, in large measure, by the Armed Forces Institute of Pathology since most of the biomedical personnel were on the AFIP staff. Over and above the contribution made by the AFIP, the total cumulative (estimated) cost borne by Contract #DA 49-092-ARO-130, was $167,077.25. A major portion of these (ARPA) contract funds was used to provide computer-science/technology support, principally by subcontract with the Planning Research Corporation. The role of PRC has been discussed under Project plans, and will not be considered further here.

A detailed financial statement has been prepared and submitted to the appropriate persons of ARO and ARPA.

Forward

The work described here is the product of a remarkable interdisciplinary effort, involving medicine (human and veterinary), geography, geology, cartography and computer science technology. I am pleased to have played a part in this important study, even though my role was a very small one.

Dr. Hopps and his associates are to be congratulated on their accomplishment so clearly described in this monograph. They have taken a giant step forward in adapting the technologic advances of information theory and computer science to the study of disease ecology. Their work clearly and dramatically demonstrates the feasibility of "computerized mapping of disease-environmental data", and they have produced the blue prints for an effective system. Hopefully, this work will continue, and the system which they have designed with such care will be implemented so that the many scientists who are interested in the relationships between the many elements of man's environment and his variety of diseases can work more effectively for the benefit of mankind throughout the world.

HARLAN I. FIRMINGER, M.D.

Preface

The truth is rarely pure and never simple.

Oscar Wilde

Communication means the perspicuous transmission of information. As with all things, the means has to be appropriate to the end. This has been an important consideration in the MOD project. Our emphasis on maps as a means of display is because this is the most effective way to transmit disease/environmental information to most persons. Often a map is the best means of presenting interrelationships vividly and dramatically -- especially to the intelligent, concerned, non-medically, non-mathematically oriented person who needs to know.

Maps allow quick and clear correlation and serve a very important need, even for the medical expert. Overlaying and visual pattern comparing is a very powerful process because it permits human detection of relationships so complex that standard mathematical methods may be unable to detect them. The rapid production of maps by computer gives an additional great advantage; the process is so fast that one can get an up-to-date presentation several hours after his request.

The computer techniques which allow map print out also allow print out of figures and names to provide specific yes/no/or qualified answers, lists of references, location of things by latitude/longitude coordinates, political areas, etc., correlation coefficients, data for construction of graphs, etc. etc.

Preface

Simply stated, the objective of the MOD system is to provide a means whereby the DISEASE PANORAMA can be quickly and effectively presented in map form in a time context which is either current or historic.

> *We mean disease panorama to include location of the disease at a particular time in terms of prevalence, incidence, mortality, and morbidity -- within the population en toto, also its various segments. But more than this: we mean it to include also information as to the quality, quantity, and location of those numerous environmental factors which influence rate of occurrence as well as character of the disease.*

In a recent (unsigned) article on the role of computers, it was said: "When all the pertnent facts are known, decisions make themselves." This may well be true, but it assumes that the pertinent facts are distinct from the great mass of non-pertinent facts, and that the degree of their pertinence is recognized. This brings us to one of the critical points in today's "information explosion". Data is increasing at an exponential rate and is inundating us because of its enormous volume. We have by no means solved the problem of converting data to information.

One of the very important problems in Geographic Pathology is how to handle the great accumulation of knowledge so that what is known shall be available when needed -- and we are talking about information, in contrast to isolated data. Geographic Pathology is not alone in facing this problem, but it is particularly serious here because the pertinent information is more widely scattered, and a higher proportion of it appears in the form of conference proceedings, annual reports from isolated medical centers, and the like. In fact, much very valuable information (stemming from actual field experience) is not written down at all, but is nevertheless available if the proper approaches are used. An inherent part of this

data processing problem has to do with separating the pertinent from the non-pertinent, differentiating between the true information and the mis-information (which abounds), all as a prelude to converting isolated facts to correlated INFORMATION. Although electronic data processors are very efficient at storage and retrieval of data, and can carry out very complex searches for correlates, these machines cannot do the essential selection and preprocessing of data which, among many other things, includes a value judgement as to the validity of the data which is fed to them.

In the course of our work, critics have pointed out to us the inherent limitations of many of the data that we would like to process. We recognize this fact very well indeed. Certainly there are many places in the world where the data base that deals with many disease situations is altogether inadequate for any meaningful collation, much less effective computer manipulation. *No system of information processing can convert bad data into good data.* However, there are large pools of data (derived from cultural anthropologists, economists, geologists, meteorologists, agronomists, epidemiologists, veterinarians, pathologists, etc.), relating to disease/environment situations, which could be meaningfully collated and effectively computer manipulated.

Many of the most important problems have the softest information, but we must identify what information there is, and learn its limitations. We must work toward correcting deficiencies in the data base, but even more important, we must develop better methods of using what information is available.

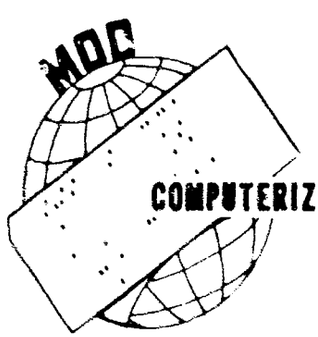
We are at the stage of world development where many important judgments must be made in the absence of hard data. If we do not use what data is available, what shall we use?

> *"Where attainable knowledge could have changed the issues, ignorance has the guilt of vice."*
>
> *Alfred North Whitehead*

Preface

The prime objective of the MOD project was to develop a system whereby *available* data could be used most effectively to gain new insight into the multifactorial causes of disease. This report is an account of how we set about to do this, the many problems that we encountered -- and our efforts to overcome these problems.

HOWARD C. HOPPS, M. D.



COMPUTERIZED MAPPING OF DISEASE

Acknowledgements

The Computerized Mapping of Disease (MOD) project is grateful for the aid it has received from many, many persons and organizations.

First, UAREP acknowledges the financial support which made this study possible: a contract from the Advanced Research Projects Agency, administered through the Army Research Office, and support through the Army Medical Research and Development Command.

Second, we express our appreciation to those persons in the Institute of Defense Analysis whose initial stimulus led to our undertaking this task, particularly Dr. Herbert Pollack, and whose subsequent guidance, through the "Mapping of Disease Advisory Committee," gave valuable assistance, also to personnel of ARO, especially Dr. Alan L. Forbes. In addition, we welcome the opportunity to extend heartfelt thanks to Major General Joe M. Blumberg who, as Director of the AFIP, gave strong encouragement and support during the early stages of the project and, later, as Commanding General, Army Medical R&D Command, continued his encouragement and support.

Third, those of us working directly on the MOD project acknowledge our deep debt of gratitude to the professional and administrative staffs of UAREP, especially Dr. R.E.Knutti and Mr. M.J.Lundfeld, and the AFIP, (including the Medical Illustration Services), especially Captain Bruce H. Smith, MC, USN, The Director. The resources of these organizations have been given unstintingly and have been invaluable to us.

Fourth, we wish to thank the various members of the many other organizations -- governmental, educational, and industrial -- from whom we have received advice and counsel, and who have influenced in ways too numerous to mention the development of our concepts and the ideas necessary to implement these concepts. A list of these organizations appears at the end of this Section. It would not be appropriate to name each of the particular

individuals at these institutions who helped us, but several of them deserve special mention: Dr. Aaron Alexander (WRAIR) who shared his great knowledge of leptospirosis with us and placed his extensive reprint file at our complete disposal; Dr. R. Warwick Armstrong (U. of Ill.) who gave valuable suggestions as to medical geographic aspects of the project (and reviewed portions of the manuscript), along with others at the University of Illinois (Center for Zoonoses Research, School of Veterinary Medicine, and Department of Geography), including Dr. Carl A. Brandly, Dr. Lyle E. Hansen, Dr. R. H. Kokernot, Dr. C.S.Alexander and Dr. Alfred W. Booth; and Dr. Paul R. Schnurrenberger of the Illinois State Department of Health.

Very special thanks are due to Mrs. Frances Wooter who typed the (photo-ready) manuscript rapidly and effectively under conditions that were much less than optimal. We acknowledge, too, the efforts of Mrs. Alfreda E. Prince, who typed most of the draft of this report, and Miss Mildred Bruce and Mrs. Betty Stone, who also helped in preparation of the manuscript. We also extend our appreciation to Don McDorman and Ron Mayor and Tom Simms and Ernest Kontor for helping to paste up and assemble the manuscript.

Organizations visited

Members of the following organizations generously provided information and advice to the MOD study team regarding various cartographic and/or biomedical data-processing procedures and data.

Alden Electronic Impulse & Recording Equipment Co., Inc.

American Geographical Society

American Institute of Biological Sciences, Biological Sciences Communication Project

American University, Dept. of Geography

Arco Corp.

Atlantic Research Corp.

Auerbach Corp.

Benson-Lehner Corp.

BioSciences Information Service of Biological Abstracts (BIOSIS)

Bowman-Gray School of Medicine, Pathology Records Retrieval Program

Bunker-Ramo Corp.

California Computer Products, Inc. (CALCOMP)

Catholic University of America, Dept. of Geography

Chemical Abstracts Service

Control Data Corp. (CDC)

Electronic Associates, Inc. (EAI)

FMA, Inc.

General Motors Corp., Allison Div.

George Washington University, Dept. of Geography

Georgetown University, School of Medicine

Geo-Space Corp.

Gerber Scientific Instrument Co.

Harvard University, Laboratory for Computer Graphics

Howard University, Dept. of Geology & Geography

Illinois Natural History Survey

Illinois State Geological Survey

Indiana University, Dept. of Astronomy, and Dept. of Geology

International Business Machines Corp. (IBM)

London School of Hygiene and Tropical Medicine,
Dept. of Parasitology

McLean Paleontological Laboratory

Rand Corp.

System Development Corp. (SDC)

Systems Research Group, Inc. (SRG)

Thailand Govt., Royal Thai Army, Medical Service

United Kingdom Govt., Ministry of Overseas Development,
Dept. of Technical Cooperation

Unites States Government:

Dept. of Agriculture, Washington Computer Center

Bureau of Census, Computer and Data-Processing Dept.

Central Intelligence Agency (CIA), Medical Division

Clearinghouse for Federal Scientific and Technical
Information (CFSTI)

Computer Sharing Exchange

Department of Defense:

Aeronautical Chart and Information Center (ACIC)

Air Force Data Services Center (AFADS)

Air Force Technical Applications Center (AFTAC)

Armed Forces Pest Control Board (AFPCB)

Army Map Service (AMS)

Army Materiel Command (AMC), Foreign Science and
Technology Section, and Systems Development
and Design Division

Army Natick Laboratories, Earth Sciences Division

Army Research Office (ARO), Scientific and Technical Information Section (STINFO)

Defense Documentation Center (DDC)

Defense Intelligence Agency (DIA)

Military Entomological Information Servi (MEIS)

Naval Command Systems Support Activity (NAVCOSSACT)

Naval Oceanographic Office (NAVOCEANO)

Naval Weapons Laboratory (NWL)

Strategy & Tactics Analysis Group (STAG)

Walter Reed Army Institute of Research (WRAIR)

Geological Survey, Branch of Geochemical Census, and Map Information Office

Library of Congress, Map Division, and National Referral Center for Science and Technology

National Aeronautics and Space Administration (NASA), Goddard Space Flight Center

National Bureau of Standards (NBS), Center for Computer Science & Technology

National Institutes of Health (NIH), Division of Computer Research & Technology, Environmental Health Division, National Cancer Institute, and National Institute of Allergy and Infectious Diseases

National Library of Medicine (NLM)

National Oceanographic Data Center (NODC)

Public Health Service (PHS), National Communicable Disease Center (NCDC)

Smithsonian Institution (Natural History Museum), Dept. of Paleobiology, and Dept. of Vertebrate Zoology

Weather Bureau

Univac Div. (of Sperry-Rand Corp.)

University of Buffalo, School of Medicine, Computer Center

University of Illinois, Dept. of Computer Science, Dept. of Forestry, Dept. of Geography, Div. of Human Ecology, School of Veterinary Medicine, and Center for Zoonoses Research

University of Kansas, State Geological Survey of Kansas

Acknowledgements

University of Maryland, Dept. of Geography, and School of Medicine

University of Michigan, Dept. of Geography

University of Missouri, College of Medicine, Medical Center
Computer Program

Woodard Research Corp.

People who are interested in data / information systems are of two general types: The first is likely to say, "I don't give a damn for your opinion; show me your data." The second, "I'm not interested in the details; I want information." The system we are describing in this report would satisfy both types of user.

"A fresh instrument serves the same purpose as foreign travel; it shows things in unusual combinations. The gain is more than a mere addition; it is a transformation."

Alfred North Whitehead

Table of contents

Forward (by Harl .. I. Firminger) i

Preface (by Howard C. Hopps) ii

Acknowledgements vi

Table of contents xiii

List of figures xvii

1. INTRODUCTION 1- 1

1.0 General considerations 1- 2

1.1 Host-parasite relationships and the ecology of disease . . . 1- 2

1.2 Objectives of the MOD project 1- 4
1.21 Definition of goals 1- 5
1.22 Maps as a means of displaying information 1- 8
1.23 Selection of diseases to study 1- 10
1.24 The data base 1- 12
1.25 Hardware/software considerations 1- 14

2. TECHNICAL SUMMARY 2- 1

2.1 Overview 2- 2

2.2 Method of approach 2- 4

3. OUTPUT ANALYSIS 3- 1

3.0 General considerations 3- 2

3.1 Types of output considered 3- 3
3.1.1 Narrative and tabular reports 3- 4
3.1.2 Graphs 3- 9
3.1.3 Maps 3- 10
3.1.4 Block diagrams 3- 11

3.2 Map considerations 3- 11
3.2.1 Categories of maps 3- 12
3.2.2 Usefulness of maps 3- 13
3.2.3 Conventional maps 3- 14
3.2.4 Disease maps 3- 16
3.2.5 Symbolic representations on maps 3- 19

3.2.5.1 Dot-type (data-point-type) maps 3- 21
3.2.5.2 Shading-type maps 3- 22
3.2.5.3 Contour-type maps 3- 29
3.2.5.4 Combination-type maps 3- 55
3.2.6 Position, scale, and projection 3- 55
3.2.7 Map construction 3- 65
3.3 Block diagrams 3- 98
3.4 Graphs 3- 99
3.5 Conclusions 3- 104

4. DATA CHARACTERISTICS 4- 1
4.0 General considerations 4- 2
4.1 Data structuring terminology 4- 3
4.2 Factor catalogue 4- 14
4.2.1 Disease factors 4- 15
4.2.2 Environmental factors 4- 27
4.3 Data for mapping -- minimal and optimal 4- 36
4.4 Problem areas related to data characteristics 4- 40
4.4.1 Data structure limitations 4- 40
4.4.2 Locations and values of data points 4- 41
4.4.3 Unreliable data 4- 45
4.4.4 Incomplete data 4- 48
4.4.5 Contradictory and erroneous data 4- 49
4.4.6 Secondary data points 4- 50
4.4.7 Location terminology 4- 51
4.5 Types and characteristics of data sources 4- 52

5. DATA COLLECTION 5- 1
5.0 General considerations 5- 2
5.1 Methods of collecting data 5- 2
5.2 Data collection activities 5- 5
5.3 Data extraction procedures 5- 6
5.4 Data input operations 5- 20

6. COMPUTER SYSTEM REQUIREMENTS 6- 1

6.0 General considerations 6- 2

6.1 Hardware requirements 6- 5
6.1.1 Output devices 6- 6
6.1.2 Input devices 6- 8
6.1.3 Storage devices 6- 9
6.1.4 Central processing units 6- 11
6.1.5 Available services 6- 11

6.2 Software requirements 6- 12
6.2.1 Available languages 6- 12
6.2.2 Available services 6- 13

6.3 Conclusions 6- 13

7. DATA PROCESSING 7- 1

7.0 General considerations 7- 2

7.1 Storage subsystem 7- 4
7.1.1 Data input cards 7- 4
7.1.2 Data file 7- 10
7.1.3 Dictionary file 7- 14
7.1.4 Dictionary input cards 7- 21
7.1.4.1 MOF construction (or reconstruction) 7- 22
7.1.4.2 Dictionary updating 7- 24
7.1.4.3 Dictionary correction 7- 27

7.1.5 Storage processing 7- 30
7.1.5.1 Dictionary building and maintenance 7- 31
7.1.5.2 Data file processing 7- 33

7.2 Retrieval subsystem 7- 37
7.2.1 Retrieval language 7- 38
7.2.2 Retrieval request cards 7- 43
7.2.3 Retrieval processing 7- 43
7.2.4 Alternate LOF coding procedure 7- 47

7.3 Synthesis subsystem 7- 51
7.3.1 Dictionary file (location functions) 7- 51
7.3.2 Query requests 7- 61
7.3.3 Calculations 7- 62
7.3.4 Sorting 7- 66
7.3.5 Combinations 7- 68
7.3.6 Enhancement 7- 71

7.4 Output subsystem 7- 73
7.4.1 Reports 7- 73
7.4.2 Maps 7- 75
7.4.2.1 Projection 7- 76
7.4.2.2 Gridding 7- 76
7.4.2.3 Production of maps 7- 77
7.4.3 Multiple output 7- 80

8. OUTPUT USAGE 8- 1

8.0 General considerations 8- 2
8.1 Operational procedures 8- 3
8.2 Notes to user 8- 6
8.3 Potential applications 8- 9
8.4 Examples 8- 12

9. GENERAL SUMMARY, CONCLUSIONS AND RECOMMENDATIONS 9- 1
9.1 General summary 9- 2
9.2 Conclusions and Recommendations 9- 4

References cited . . . . . R- 1

Selected bibliography . . . . . B- 1

Appendix . . . . . A- 1
Glossary
-- Computer processing terms . . . . . A- 2
-- Biomedical terms . . . . . A- 6
Data sources
-- Narrative and tabular . . . . . A- 10
-- Published maps . . . . . A- 16
Schistosomiasis: general considerations . . . . . A- 25
Leptospirosis: general considerations . . . . . A- 28

List of figures

FIGURE		PAGE
1-1.	Generalized concept of the MOL system	1- 17
3-1.	Listing of the standard set of schistosomiasis data used during MOD mapping studies	3- 5
3-2.	Two- and three-variable graphs illustrating possible disease/environmental relationships	3- 6,7,8
3-3.	Distribution of Burkitt's tumor (in Africa) related to environment	3- 17
3-4.	Distribution of goiter (in U.S.) related to low iodine content of drinking water	3- 18
3-5.	Flow-line-type map illustrating spread of "Asian flu" over the world	3- 20
3-6.	Manually drawn, dot-type map showing the amount of land in farms and the percentage of that land available for crops	3- 23
3-7.	Examples of machine-drawn, dot-type maps	3- 24
3-8.	Manually drawn, dot-type map depicting various leptospiral serotypes in different countries	3- 25
3-9.	Dot-type maps showing South American schistosomiasis data, manually drawn as part of the MOD study	3- 26,27
3-10.	Dot-type map prepared from the South American schistosomiasis data, using a trend-surface program	3- 28
3-11.	Manually-drawn, shading-type map showing types of natural vegetation	3- 30
3-12.	Examples of machine-drawn, shading-type maps	3- 31
3-13.	Manually drawn, shading-type map showing infant mortality (in Great Britain)	3- 32

3-14. Manually drawn, shading-type map prepared from the standard set of South American schistosomiasis data 3- 33

3-15. Same data expressed in a shading-type map, produced by a computer/line-printer 3- 34

3-16. Relationship between an actual land surface and its representation as a contour-type map 3- 37

3-17. Land surface (and data points) used to demonstrate construction of a contour-type map 3- 39

3-18. Data points plotted on a square (LO, LA) grid 3- 40

3-19. Interpolation of contour-line values around two data points 3- 41

3-20. Interpolated data points with VAL = 70 feet, 60 feet, and 30 feet 3- 43

3-21. Preliminary 70-, 60-, and 30-foot contour lines 3- 44

3-22. Final 70-, 60-, and 30-foot contour lines 3- 45

3-23. Final contour lines, drawn from the data points of figure 3-17 B 3- 46

3-24. Illustrations of rules to be followed when drawing contour lines 3- 48

3-25. Manually drawn, contour-type map showing the environmental factor, "mean sea-level temperature in July" 3- 50

3-26. Examples of machine-drawn, contour-type maps 3- 51

3-27. Manually drawn, contour-type map depicting mortality rates from arteriosclerotic and degenerative heart disease (in Australia) 3- 52

3-28. Contour-type map of standard MOD set of schistosomiasis data drawn manually 3- 53

3-29. Manually drawn contour map based on same data, computer manipulated 3- 54

3-30. Manually drawn map combining contour-type and shading-type symbols to show population data 3- 56

3-31. Machine-drawn, combination-type maps 3- 57

3-32. Manually drawn map using a combination of dot-, shading-, and contour-type symbols to portray urban infant mortality in India 3- 58

List of Figures

3-33. The standard MOD set of schistosomiasis data presented as a manually-drawn map, using dot-, shading-, and contour-mapping techniques 3- 59

3-34. The same MOD data presented as a combined dot-type and contour-type map, produced by a computer/plotter 3- 60

3-35. Various map projections most useful to the MOD system 3- 62,63

3-36. MOD (Venezuelan) schistosomiasis data: shading-type maps produced to simulate computer/line-printer and computer/plotter output 3- 67

3-37. Shading patterns combined within a single grid box 3- 68

3-38. Shading-type map, presenting MOD schistosomiasis data as produced on a line-printer 3- 70

3-39. Manually drawn, contour-type maps showing African schistosomiasis data 3- 72,73

3-40. The standard MOD South American schistosomiasis data mapped manually, using a three-point-plane method 3- 75

3-41. MOD (Venezuelan) schistosomiasis data mapped to illustrate the effects of variable grid sizes on contouring operations 3- 76

3-42. The standard schistosomiasis test data, contour-mapped by Control Data Corporation's contouring system, using a 1° grid 3- 78

3-43. The same data, contoured by Control Data Corporation's program, using a 2° grid 3- 79

3-44. Rabies data, contoured by Control Data Corporation's system, also manually 3- 81,82

3-45. The standard MOD South American schistosomiasis data, contour-mapped by the University of Michigan's contouring program 3- 83

3-46. Computer/line-printer produced map from which were taken the contour lines of figure 3-45 . . . . . 3- 84

3-47. Standard MOD schistosomiasis data, contour-mapped as sixth-degree trend surface 3- 86

3-48. Computer/line-printer produced map from which the contour lines of figure 3-47 were drawn 3- 87

3-49. MOD (Venezuelan) schistosomiasis data contour-mapped, using different methods of grouping data 3- 89

3-50. Line-printer drawn map showing (by numbers) the known data points of figure 3-18 and (by ?'s) the unknown data/grid points in between 3- 89

3-51. MOD schistosomiasis data for eastern Brazil, contoured by spreading the reported data over all of each reporting area 3- 91,93

3-52. Schistosomiasis in eastern Brazil: a MOD-produced map and a published map showing comparable data 3- 94

3-53. MOD schistosomiasis test data (for eastern Brazil), machine-mapped using the Naval Oceanographic Office's program 3- 95

3-54. Manually produced topographic map compared with a computer simulated map prepared from the same (selected) data 3- 97

3-55. Manually drawn block diagram portraying the form of the land's surface 3- 100

3-56. Computer produced block diagram showing the distribution of the human population in the United States 3- 100

3-57. Standard set of schistosomiasis data presented as a block diagram, drawn manually 3- 101

3-58. Same set of data shown as a computer produced block diagram. 3- 102

4-1. Map made from illustrative data points (discussed in text) to show function of various portions of the (MOD) structured data 4- 13

4-2. Analogy between MOD data structure and an orchard 4- 35

5-1. A suggested pattern of data collection 5- 4

5-2. MOD form for extracting minimal (for mapping) leptospirosis data 5- 10

5-3. MOD form for extracting more than merely minimal (for mapping) leptospirosis data 5- 11,12

5-4. MOD form for extracting standard schistosomiasis test data 5- 13

List of Figures

5-5. MOD form for extracting rabies data 5- 14

5-6. MOD extraction form for data relating to biogeographic distribution of small mammals 5- 15

5-7. MOD data collection form used in compiling file of published environmental-factor maps of southeast Asia 5- 16

5-8. Basic procedures followed in entering data into the MOD computer system 5- 21

7-1. MOD computer system flow diagram 7- 5

7-2. Manner in which preprinted LOF code numbers function in automated validity checking 7- 6

7-3. MOD Storage Subsystem flow diagram 7- 30

7-4. MOD Retrieval Subsystem flow diagram 7- 39

7-5. MOD Synthesis Subsystem flow diagram 7- 53

7-6. MOD Output Subsystem flow diagram 7- 72

8-1. Overall pattern of MOD system usage 8- 4,5

8-2. Types of relationships among disease and environmental data 8- 8

8-3. Data concerning Burkitt's tumor recast into MOD-like output form 8- 14

8-4. Data implying relationship between goiter and iodine content of drinking water, rearranged into MOD-like format 8- 15

8-5. Maps showing distribution of temperature, rainfall, and schistosomiasis in eastern Brazil 8- 16,17

8-6. MOD-type maps, each showing the distribution of one environmental factor pertinent to the study of a paleontologic problem 8- 19

8-7. Overlaid combinations of appropriate MOD-like maps from figure 8-6 to show use of maps in resolving the paleontologic problem 8- 21

9-1. Relative amounts of effort necessary to operate the fully implemented MOD system 9- 5

9-2. Suggested table of organization for personnel involved in operating the (implemented) MOD system 9- 7

A-1. A schematic presentation of the life cycle of schistosomiasis . . . . . A- 27

A-2. A schema showing some aspects of the ecology of leptospirosis . . . . . A- 30

Separated, experimental data and medical judgement scarcely lend each other assistance; associated in the same intellect, they illumine and enrich one another.

Ramon Y Cajal

1 Introduction

ABSTRACT - This section discusses multifactorial causes of infectious disease and develops the concept of disease ecology. Against this background, the objectives of the MOD project are presented, and the ways and means of realizing these objectives. Particular advantages of the map-form, as a means of displaying information, are discussed.

One cannot study man alone in relation to his diseases because here -- as almost everywhere else -- man is inextricably bound to his environment.

1.0 GENERAL CONSIDERATIONS

We are faced with the problem that always arises when one addresses a mixed audience. Presumably this audience (those who read this book) will consist of experts, but some will be experts in the field of medicine, or geography, or political economy, and know relatively little about data processing or computer technology; others, highly skilled in various aspects of information theory and communication science, will know relatively little about medical geography or the biodynamics of disease.

We have tried to reach an acceptable compromise by laying a groundwork of basic information at the beginning of each section before commencing the technical discussion. To familiarize some of the non-medically oriented readers with schistosomiasis and leptospirosis, a brief discussion of each of these diseases is given in the appendix. Since a common ground of understanding is dependent upon knowing precisely what the writer means when he uses certain words, we have provided a Glossary (also in the Appendix), and this is divided into two parts: Part 1 dealing with computer processing terms, Part 2 dealing with biomedical terms.

1.1 HOST-PARASITE RELATIONSHIPS AND THE ECOLOGY OF DISEASE

In a general sense, the MOD project is primarily concerned with just this: host-parasite* relationships. But there is a vast arena in which these relationships develop, an arena which contributes in a very important way to these relationships. Thus we must consider, but go well beyond

---

*The term parasite is used here in its broadest sense to include all living agents that live in or on a "host", deriving benefit from the host -- and these agents include viruses, bacteria, spirochetes, yeasts and fungi, as well as parasitic agents (in the narrow sense).

factors which directly concern just the host or just the parasite. Some of these additional factors affect the interface between the potential host and potential parasite (relating to the critical act of infection, per se), others affect the host's response to the parasite -- and vice versa -- (relating to the disease, per se).

Factors primarily concerned with the host have been described as follows (Hopps, H.C., Principles of Pathology, p. 388, 2nd ed., Appleton, Century, Crofts, New York, 1964).

> *Differences among individuals are, of course, very important in determining the diseases to which they are susceptible, and their reactions to the diseases once they contract them. But patterns of disease, involving large groups of people, is a very different matter, and provides a quite new perspective in our study of disease. We can learn much of value by looking into the reason for these varied patterns of disease. The principal factors include: (1) Time, in world history, (2) Age, i.e., time in the life of the individual, (3) Race, (4) Sex, (5) Socio-economic conditions and customs, and (6) Geographic location.*

Factors primarily concerned with the parasite are more numerous and more complex because, in addition to involving many physical and chemical aspects of the environment, the parasites may be dependent upon intermediate hosts and/or insect vectors to complete their life cycle, and may also be dependent upon animal reservoirs as a means (either direct or indirect) of reaching their definitive host. (For those who are not biomedically oriented, it may be advantageous at this point to get some basic orientation by reading the brief discussions of leptospirosis and schistosomiasis that are to be found in the Appendix.)

Taking the host, the disease agent, and THE ENVIRONMENT all together, one has a complex relationship that is properly termed the ecology of disease.

An understanding of disease ecology is essential if one is to comprehend the reasons why disease is such a varying entity -- why the "same" disease (in terms of its etiologic agent) can affect an occasional person under some conditions and whole populations under others; why sometimes it may be so mild as to escape detection and other times rapidly fatal.

Folke Henschen's description of the dynamic character of disease is quite appropriate to this discussion (The History and Geography of Diseases, p. I ((Introduction)), Delacorte Press, New York, 1966).

> *Diseases are not unchanging phenomena. Their appearance and character are subject to historical development and varying geographical and demographical conditions of population. Some diseases seem able to disappear; other new ones to appear. Infectious diseases, which only one or two generations ago formed the largest group in our statistics on morbidity and mortality, have been driven back by the advance of medicine. Instead two other groups of diseases, cardio-vascular diseases and tumours, have taken the first place, a development which is partly connected with a rising average expectation of life. However, this revolution affects, above all, North America and many of the countries of Europe. But even in those countries whose populations form the great majority of the world's inhabitants, a development in the same direction has occurred. The overall picture of diseases within one country or community, which one can call the 'disease-panorama', varies then from time to time, from country to country, and from town to town.*

1.2 OBJECTIVES OF THE MOD PROJECT

Objectives are discussed under five subheadings, beginning with a definition of goals.

1.2.1 DEFINITION OF GOALS

At the 37th session of the Executive Board of the World Health Organization (January 1966) Professor Murray Eden gave an important "Statement of Communications Science" in the course of which he made the following assertions:

- *That medicine and biology are concerned with the observation of patterns which in principle can be made precise only with the help of mathematics.*
- *That computers and computer science can open up an entirely new way of studying health problems.*
- *. . . that there is a language barrier between the communications scientists on the one side and the physicians, biologists and health administrators on the other; a barrier which must be breached if these new techniques are to work together with medical science (and they must work together) for the betterment of the health of people.*

The MOD project was undertaken (in 1964) with precisely these ideas in mind.

The Computerized Mapping of Disease Project (MOD) has two principal objectives. The first, and most important, is <u>to develop a system that will provide for</u>:

(1) Recording, classifying, collating, and validating a wide variety of medical-environmental data.

(2) *Preprocessing* the medical-environmental data so that it can be computer processed.*

---

* One of the major problems is to structure a data analysis vocabulary, developing a heirarchial system for the qualitative and quantitative characterization of disease/ecologic information. This requires cutting across disciplinary boundaries, identifying the "common denominator" of the various jargons, and converting the narrative and tabular data into a miscible form. (This is quite different from developing a dictionary or thesaurus.)

(3) Development of a storage/retrieval mechanism to act upon such preprocessed medical data, together with a complex editing program that will allow updating, that will provide for immediate identification of material in conflict, and that will print out specific data sources.

(4) Development of programs that will allow manipulation of the data to show significant interrelationships.

(5) Development of programs whereby the computer can "instruct" a plotter to prepare contour maps reflecting quantitative aspects of incidence/prevalence of specific diseases, together with distribution of a wide variety of causally related factors, e.g., climatic factors, soil factors, animal reservoirs and/or insect vectors, characteristics of the human population, etc., etc.

(6) Development of programs whereby supporting information (to accompany the maps) can be printed out, extending the usefulness of the mapped medical information.

(7) Development of programs whereby other types of graphic display can be generated to show cause/effect relationships (e.g., line graphs) pertaining to prevalence and/or incidence of a given disease.

The second objective of the MOD project is to produce meaningful maps (and other graphic displays) that show the distribution of a disease(s) in terms of prevalence, incidence, severity, etc., along with distributions of selected causally related factors. Quantitative as well as qualitative aspects will be considered, with major emphasis on contour-type maps, the contour lines representing isarithms.

1. *Introduction*

By using a computerized system of analysis and output, it will be possible to produce distribution maps in a matter of minutes rather than months, as has previously been the case. This will allow up-dating whenever required. Furthermore, such a system will permit the production of many more maps than would otherwise be practical, covering a wide range of ecologic factors. As desired, these could be printed on transparent stock suitable for overlay assembly in order to compare one pattern of distribution with another, etc.* In the past, the time involved in preparing disease distribution maps has been prohibitive in terms of maintaining current information. This is reflected by the fact that, to date, there have been only two major contributions in this field:

> A Geographic Atlas of Disease prepared by the American Geographical Society and published during 1950-55.
>
> A World-Atlas of Epidemic Diseases edited by Professor Ernst Rodenwaldt (Heidelberg) and published in 1952 (but reflecting data gathered some years before).

From a broader point of view, the MOD Project (Mapping Of Disease) is an effort to illuminate the *geographic pathology* of disease. Geographic pathology is, in a sense, a kind of comparative pathology -- one in which place (rather than species) is the primary variable. Geographic pathology attempts to answer the questions: What (disease); Where (is it) -- and When; and Why (is it there). Of course geographic pathology includes aspects of epidemiology since it, also, is concerned with prevalence and incidence and the interplay among complex causal factors, but it goes beyond epidemiology in its concern for the pathogenesis and the pathologic effects of the disease under study.

---

* There is virtually no computer limitation of map scale; as the geographic area to be covered decreases (the size of the map remaining constant) the map scale varies inversely and "resolution" increases.

Potential Applications of the MOD system, in addition to producing disease distribution maps, per se (and other graphic displays), and in helping to determine causal relationships, include: (a) use in evaluating probable (disease) consequences from particular changes in ecology, and (b) use in developing mathematical models by which one may predict major changes in disease incidence, e.g., epidemics.

Although the MOD system has been developed with primary concern for infectious diseases, the system is applicable to virtually any problem area in which "things" need be considered in a time/location context.

> Summarizing, the MOD project is an effort to: (1) characterize input data (relating to disease/environment) in such a way that they can be stored and readily retrieved *in context* by a computerized system which, (2) using these data, can relate meaningfully, prevalence/incidence/character of disease to a variety of direct and indirect causal factors, and (3) output the information directly in map form.

1.2.2 MAPS AS A MEANS OF DISPLAYING INFORMATION

Maps were chosen as the principal pattern-form to display information pertaining to disease (but not the only means) for two reasons: First, because those areas in which the distribution of disease agent and host overlap mark the geographic regions where the disease can occur; evaluation of such ecologic factors as temperature, rainfall, humidity, the amount and mineral content and pH of surface water, agricultural practices, population densities of various plants and animals (including man), the kinds of people involved (not only age and sex, but race, ethnic group, and tribe) and their customs -- and a hundred other factors closely tied to geographic location -- can help us to determine where the disease will occur, and how it will be

manifest. Second, because maps have a unique advantage over most other forms of graphic display for the general reasons that:

(1) Extensive and continual usage of map forms, beginning in early childhood, has conditioned most (educated) people to an intuitive understanding of maps.

(2) The map is ideally suited to a consideration of multiple factors simultaneously (e.g., place -- both geographic and political -- in relation to topography, population density, the location of towns and cities, the location and character of transportation routes, and time zones).

(3) Through the use of rather simple devices such as isarithms (isopleths) one can achieve a three dimensional effect in a two dimensional presentation (quantity becomes the third dimension; quality and location the other two).

We believe that a mechanism/system which can produce many kinds of map-patterns quickly, in response to specific query, will offer two very important advantages: First, such a mechanism will make it possible to have current information about the distribution of specific diseases and the distribution of *known* causally related agents or conditions. Second, the rapid availability of a large number and wide variety of disease-environmental maps will give the observer an opportunity to compare location patterns of *unknown* but possibly related ecologic factors and, in this way, help him to identify causal relationships that might otherwise have escaped notice.

Two recent articles in Nature describe the present situation well in terms of needs and accomplishments in the field of automatic data processing/computer mapping, and these comments are very pertinent to the MOD project.

> *25 March 1967 --". . . only a tiny proportion of the mass of demographic and climatic information collected by governments ever sees print in map form. Information is simply tabulated by area, and the possibility of spotting regularities or correlations -- say the incidence of pellagra, family expenditure on food, and the provision of medical services in the eastern United States -- is very remote indeed. Such interesting relations as have been found are the result of years of searching, and merely increase the sense of frustration that there is no better way of doing it."*

18 April 1967 -- "In the environmental sciences, a shortage of information has been turned into a flood by advances such as automatic data loggers, photographic survey and satellite instrumentation, but methods of using this information have been left behind. Maps are the best way of making the environmental information comprehensible, but they are still prepared by a slow, expensive and inflexible process. Professor Linton of the Department of Geography at the University of Birmingham described the process bluntly as "cottage industry". So far, unfortunately, it has not proved possible to automate the process. Automation is attractive because even if it did not make the process simpler, it would certainly speed it up. In addition to being a great advantage in conventional topographic mapping, the increase in speed would encourage people to include a locational element in observations which are not at present mapped at all. These observations, such as the census returns, could then be made available in map form as well as in statistical form.

Maps also offer a powerful research tool which many people believe is under-exploited because of the difficulties and expense in preparing maps by hand; maps showing the spread of diseases can be compared with those showing diet, economic circumstances, or the availability of medical services, and reveal unexpected correlations -- lead in the soil and the incidence of mental disease, for example.

1.2.3 SELECTION OF DISEASES TO STUDY

The major objective of the MOD project is to develop a method that will give new insight into disease-environmental relationships, but this cannot be done effectively without reference to real life situations. Disease models are necessary, not only to develop methods, but to test them. Our choice of diseases to study was governed by the following considerations:

(1) There must be adequate information about the disease in terms of amount* and qualitative/quantitative factors, and it must be possible to connect both of these items to geographic loci.

(2) The disease must be one in which geographic distribution reflects ecologic requirements.

(3) The disease should be one in which the prevalence and/or incidence is related to certain variables in a way that is understood (at least partially) -- and the variables must be measurable and of known geographic distribution.†

Selection of the disease area to study was influenced by these considerations:

(1) The area should be fairly large (to test the graphic generator programs).

(2) There should be several hierarchal levels of political units in the area (e.g., country, state and county).

(3) Disease-environmental data should be rather uniformly distributed over the area, i.e., there should be few "unknown" or "unreported" subsections within the area.

Of many many possible diseases for study, based upon the considerations given, we have concentrated upon leptospirosis and, to less extent, schistosomiasis and rabies. These were carefully chosen for several reasons:

---

* The disease must be positively diagnosable and reported. Furthermore, effects of the disease should persist for a long while unless the disease is readily apparent in its acute form and continually searched for.

† Man to man transmitted diseases should be avoided since occurrence is primarily influenced by intimacy of contact and "immunity". This excludes most upper respiratory infections.

(a) they are important diseases; (b) we (in the Geographic Pathology Division) know a good deal about them; (c) highly reliable laboratory diagnosis is possible; (d) they are wide-spread in distribution, but not diffuse; (e) more reliable distribution maps are badly needed; (f) each of the diseases poses specific data processing challenges in relation to important ecologic factors. For example:

Leptospirosis:
- Involves many mammalian reservoirs (100±), both domestic and wild.
- Occurs throughout most of the world.
- Prevalence is greatly influenced by the amount and nature of surface water in the area, including pH, mineral content, rate of evaporation, etc.
- Prevalence is greatly influenced by occupational and/or recreational habits of human beings.
- Severity varies markedly, depending upon serotype and many other factors.

1.2.4 THE DATA BASE

There are three basic parts to the MOD system and these intimately relate to each other in the sequence shown below.

Locate and get data sources

↳ Select/extract/format to produce a *DATA BASE FILE*

↳ Design/implement Software/hardware system

(giving special consideration to form of output)

The data base is, obviously, an essential (key) ingredient of the System since it provides the substance upon which the software/hardware components act.

But (as shown previously) there are two aspects to the data base: (a) collecting the raw data and (b) preprocessing this so that it can be effectively stored, retrieved, manipulated, and output as information. The difficulties in getting the raw data are very considerable, and should not be minimized. But the greatest problem that the bio-medically oriented person encounters when he attempts to use a computerized data processing system is in preparing (preprocessing) the medical-environmental data for computer input.

This problem arises because the bio-medical terms are not direct and simple; they do not have clearly defined single quality/quantity characteristics. This preprocessing phase of the procedure is aptly described by the word translation -- if one understands that we are not talking about translating a foreign language into English, or one machine language into another, etc. Translation, in this particular sense, means the reduction of complex data terms to bits which have a common denominator, so to speak, and are thus compatible with all the other bits of data which bear relationship. One can combine oranges, and grapefruit, and pineapple, and coconut -- and get ambrosia -- but there is a limit. One cannot combine, meaningfully, oranges, and minutes, and square miles, and altitude, as such. But, given a proper data structuring method, one can combine, meaningfully, orange trees, time (in a calendar sense), square miles (in a location sense), and altitude -- along with temperature, rainfall, human population (the availability of agricultural laborers, the numbers of people who drink orange juice, etc.) and the prevalence and incidence of carotenemia -- and produce distribution maps which show important interrelationships among these items.

A simpler, and more common problem in translation comes when we must make compatible two statements such as: (1) "In the early winter of 1964 there occurred a mild epidemic of influenza among the urban population of northwestern United States," and (2) "During the period 6-29 November 1964,

the prevalence of type-A influenza infection was 47% greater than for a similar period in 1963 among the inhabitants of cities with populations over 35,000 in Washington, Oregon, Idaho, and western Montana." Both statements are, obviously, very closely related, and we have no difficulty, subconsciously, in fitting them together. But without a precise system for translation, these two data-complexes would be "considered" by a computer to be entirely unrelated.

The problems of preprocessing data are further complicated in connection with the MOD system because, as we have mentioned before, any broad consideration of disease-environmental relationships must utilize data drawn from a group of different disciplines each of which, in a sense, has its own scientific language -- geography, geology, agronomy, political economy, cultural anthropology, pathology, etc. etc., but this aspect of the problem, with other aspects, is treated extensively in Section 4., Data Characteristics.

1.2.5 HARDWARE/SOFTWARE CONSIDERATIONS

Speaking about multifactorial etiology of disease, Professor A. Payne in his "Statement on Epidemiology" to the Executive Board of WHO (20 January 1966) said:

> *"These changed concepts and increased complexities require the development of new theoretical and analytical approaches. We can no longer be content with the solution of simple situations such as one agent of known infectivity, incubation period, etc., in a population of known density and immune status. Mathematical models, which involve the translation of real world problems into symbols and numbers, already exist which enable us to predict, within reasonable limits, the outcome of the introduction of an agent into such a situation."*
>
> *"The new concepts demand the formulation of models many times more complex and require both highly sophisticated mathematical treatment and advanced computer technology. Complex data of this kind cannot be handled in any other way, and formulation of new models demands the aid of mathematicians and computer scientists."*

We agree entirely with Professor Payne, and it was this point of view that led us to undertake the MOD project nearly three years ago.

The MOD effort has been directed by disease-oriented rather than computer-oriented persons because, from the outset, we realized that computer processing was the means to an end. We have worked diligently so that the criticism leveled by Dr. A Feinstein at (some) computer technologists would not be appropriate for us: "They may understand the machine, but not the problem; mathematical theory, but not the nature of the problem -- the statistics may be excellent, but irrelevant." These statements are not meant to demean the role of the computer nor the systems analysts, programers, etc. Sections 6 and 7 of this volume give clear evidence that we do appreciate the importance (and complexity) of computer technology.

Continuing with our very general consideration of computer processing it is appropriate to point out that automation does not make a process simpler, it simply speeds it up. From a practical viewpoint, however, the great speed of operation allows manipulation of a volume of data that would otherwise be virtually impossible to handle. In relation to the MOD system's (ultimate) requirements we have estimated that on the order of 10^{30} possible factors may need to be considered (not necessarily used)!

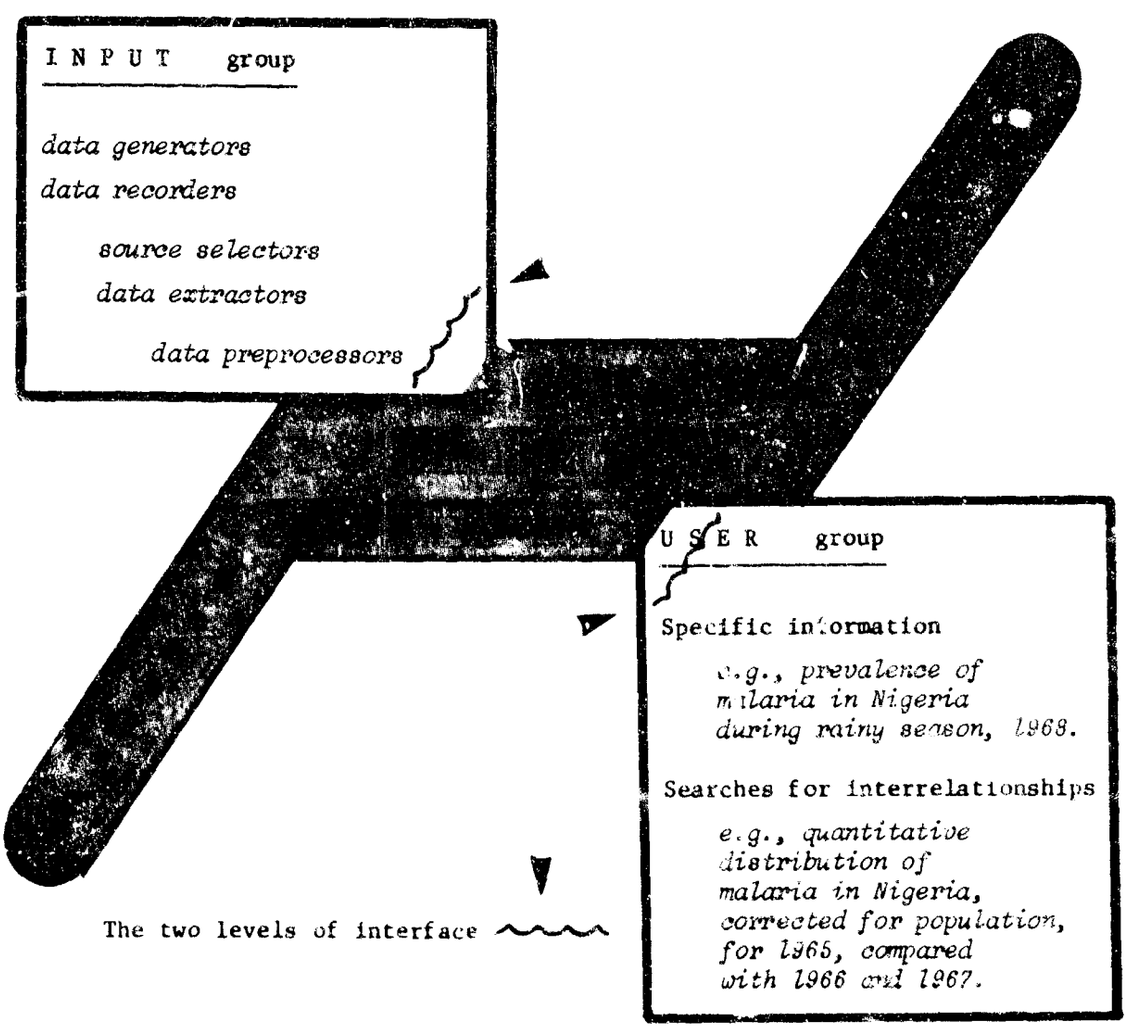
In addition to speed of operation, automation has another very important asset. It provides for a consistency of handling data that, otherwise, would not be attained. In turn, this consistency of handling forces a clear and sharp characterization of the data input, the query, and the information output.

There are many conventional aspects to the hardware/software requirements of the MOD system, but there are two unconventional aspects. The first relates to the storage/retrieval and manipulation of uniquely structured input data that deal with medical-environmental situations -- processes which include a complex Dictionary File that recognizes errors and allows

for their correction, that provides for updating, and that also peforms a gazeteer function. The second has to do with developing computer programs to contour-plot disease-environmental data -- data that are often represented by relatively sparse data points.

* * *

In a sense, the hardware/software is the computer, and the computer is the milieu in which the raw (input) material is converted to the finished (output) product. The figure below illustrates this relationship and also points up the fact that one important *interface* exists between the input group and the computer, and another between the user group and the computer.



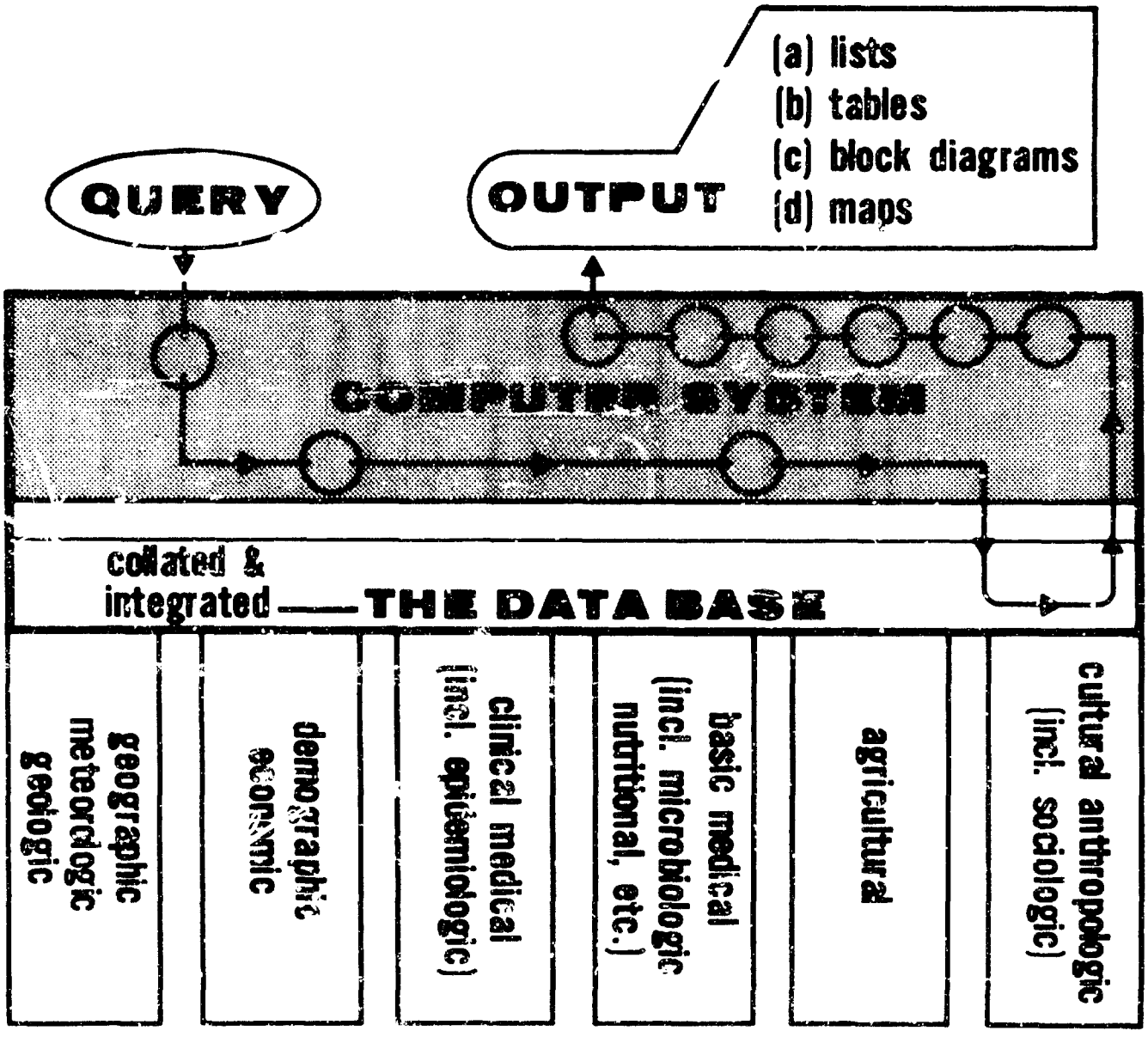


Figure 1-1 Generalized MCD system concept showing how a query is acted upon by the computer system which, first, dips into the data base to obtain data, then processes those data and, finally, outputs them as information for the user.

2 Technical summary

ABSTRACT - This section is a general consideration of what has been accomplished during the approximately 30 months that the MOD project has been active, supplementing and complementing the (prospective) view presented in the Introduction, and the General Summary, Conclusions, and Recommendations, which comprise Section 9. In addition, it explains the organization of this book.

In every problem situation we are faced with the twin questions:

- What do we know?
- What do we not know?

When we answer these questions, <u>then</u> we are ready to make progress in understanding; when we have gained understanding, <u>then</u> we are prepared to make decisions.

2.1 OVERVIEW

Section 1, the Introduction, considers the MOD project as a whole and summarizes many aspects. Furthermore, each section contains its own abstract. These are reasons why this technical summary is quite brief.

The Geographic Distribution of Infectious Disease Project is characterized by its unique conceptual framework and the variety of scientific disciplines it represents: geography, cartography, geology, meteorology, agronomy, biology, medicine, political economy, cultural anthropology, systems analysis, computer systems design, computer programing, etc.

Because of this multidisciplinary involvement, certain fundamental questions concerning the *output desired* from the system and the potential users of that product had to be resolved before beginning system design.

Useful types of output were determined to be in the form of maps, graphs, and narrative reports. Maps were selected as the principal device to display areal relationships, and were investigated in detail and defined: both conventional maps and special maps, suited to the study of geographic distribution of disease-environmental data.

After determining the output desired from a geographic disease-environmental data system, *input data characteristics* were considered. It was necessary to develop data-structuring terminology before further progress could be made, and this was accomplished. MOD's data requirements were then compiled in the form of a catalogue of disease and environmental factors, and both minimal and ideal data needs were outlined. Data sources

---

* The term, MOD, used throughout this report is an acronym derived from the initial letters of the three words that comprise the abbreviated name of the project: Mapping of Disease.

were considered and the many problems posed by particular characteristics of data available for use in the MOD system were analyzed

The combination of input and output requirements for the MOD system dictated *computerized techniques* for solution. In considering these techniques, the requirements for output devices, input devices, and central processing units were determined. Output devices are a primary consideration since the system design is based upon the required output. Cathode-ray-tube (CRT) devices, digital plotters, and line-printers were all investigated in terms of MOD system use. Although input devices are a secondary consideration, optical character recognition (OCR) machines, punched paper tape readers, card (or tape card image) readers, and digitizers were all analyzed with respect to possible use in the system. Data processing methods must be considered in any discussion of computer systems, and the basic data processing techniques and commonly-used languages were investigated. Detailed conclusions relating to the integration of computer equipment and data processing methods are presented in the appropriate sections (especially Section 6).

The system design specifications for the MOD system have been prepared in varying levels of detail. Included in this report (Section 7) are discussions of the four subsystems comprising the MOD system: storage, retrieval, synthesis, and output.

* * *

Throughout this report many figures are presented to illustrate problems encountered in designing the MOD system -- and our effectiveness in overcoming these problems. Actions speak louder than words, and we believe that the computer simulated manually drawn maps and the computer/line-printer and /plotter produced maps showing disease-environmental data speak loudly in support of our conclusions that: (1) the computerized mapping

of disease-environmental data is feasible; (2) the MOD system design we have described here represents an effective "blue-print"; and (3) this system should be implemented as soon as possible because there is a widespread and pressing need for the type of data processing and output that the MOD system would produce.

2.2 METHOD OF APPROACH

The reader is due some explanation of why this report was organized as it is presented here. The method of our approach to the problems of designing the MOD system had a major influence on the way we have approached this account of our activities.

The Preface and Introduction set the stage, so to speak, by describing the objectives of the project and considering, in general, ways and means of attacking the many problems.

Since the data are of primary consideration, Data characteristics and Data collection are considered before Computer system requirements (an essential factor in system analysis) and Data processing (a description of system design).

After this background, the Section, Output usage, describes operational procedures and potential applications of the system.

Finally, a General summary, conclusions and recommendations are given.

The Appendix presents a variety of useful information to supplement that contained in the major portion of the report.

3 Output analysis

ABSTRACT - This section considers, in detail, the types of output required of the MOD system. Since maps of disease-environmental data are of major concern, the various types of maps are explained -- and how to construct them. Block diagrams and graphs are also discussed in the context of disease-environmental relationships.

"The true purpose of knowledge resides in the consequences of directed action."

John Dewey

3.0 GENERAL CONSIDERATIONS

The most critical part in developing any automated data processing system is the determination of precisely what the output (result) should be. This is necessarily so because output requirements directly influence virtually every step in development of the system. This section gives a comprehensive evaluation of those output considerations which were the basis for designing the MOD system. It considers also the other side of the coin -- input. In a sense, input is the cloth from which the garment is made; system design is the pattern to which it is cut -- and the computer is the sewing machine.

Since output is of no value unless it is put to good use, any consideration of output is sterile unless the potential user is also considered. For convenience, output usage is considered in detail in Section 8 (after the entire system has been discussed), but output usage, as we conceived it, served as a constant guide in output analysis. Obviously, the output of the MOD system is *information*, information directed primarily toward:

- Presenting quantitative aspects of disease-environmental data in relation to place and time.
- Identifying the multiple causal factors in a given disease, and their interrelationships.
- Determining interrelationships, if any, among several different diseases occurring together, e.g., schistosomiasis, iron deficiency, protein malnutrition and tuberculosis.
- Evaluating the impact of the disease upon socio-economic aspects of the area, military operations, etc., etc.
- Anticipating the effects of altered ecology on incidence and manifestation of disease.
- Predicting variations in incidence which are likely to occur in the foreseeable future -- on the basis of past history and trend analysis.

3. Output Analysis

A system containing generalized disease information could provide output that would satisfy economic- or cultural development-oriented people who were attempting to assess the influence of specific diseases upon the development of a particular country or society (e.g., AID). Such a system would also serve an industrial group planning to establish a base of operations, guiding them in immunization and other prophylactic measures, in the design of medical facilities, in the types and specific locations of houses, dormitories, etc. etc. In other words, such an input/output system would be particularly useful to decision-makers who required information about geographically oriented disease conditions over comparatively broad regions.

A system containing more detailed information, on the other hand, would be required to satisfy biomedical researchers concerned with in-depth investigations of disease-environmental situations, particularly causal relationships. Similarly, more detailed information would be useful to public health officials whose major objective was surveillance of specific diseases.

As discussed in Section 4, Data Characteristics (4.4.1), the method by which the data is structured permits it to be entered at any level of generality, and to be retrieved at any level <u>equal to or more general than</u> that at which it was entered. In other words, highly specific data can be used in a broad or general way, but not vice versa. Thus, the more detailed the data input, the more potential users could be satisfied. Because of this the MOD system, including data extraction forms, has been designed to receive and process data in its most detailed form (when available) -- detailed as to precise geographic location as well as to specific qualitative and quantitative characteristics

3.1 TYPES OF OUTPUT CONSIDERED

The information to be output by the proposed MOD system can take several forms. For purposes of this discussion these are: <u>narrative reports</u> (i.e., listings or tables), <u>graphs</u>, <u>maps</u>, and <u>block diagrams</u>.

3.1.1. NARRATIVE AND TABULAR REPORTS

Today, most computer output takes the form of "hard-copy" reports, i.e., printed words and numbers arranged in lists, tables, or narrative-like prose. The techniques for producing these are well known and need not be discussed in detail here. It is important to realize, however, that a computer system cannot, ordinarily, combine data stored in free prose or narrative form and produce summaries of such data; but it can summarize rather rigidly formatted data and produce meaningful short reports. Although the MOD system concentrates on output in the form of maps, narrative and tabular reports are also an important output product because they are required to display such items as input data, the contents of a data file, data retrieved by queries, data to be used in generating other forms of output, etc.

Figure 3-1 shows a set of data that was extracted from Malek (in May, 1961), and illustrates one kind of tabular output useful in studying a disease situation. (The project team added longitude and latitude coordinates to the data and rounded the data values to make them more easily comparable.) This set of data was used as a standard set for investigating mapping techniques, and it appears throughout this report in various forms. We emphasize that these particular data are quite limited in scope, and that their primary use has been in developing methods for various computerized outputs during design of the MOD system.

If for each geographic locality the rat population density were plotted as the X coordinate, the pH value of the surface water as the Y coordinate, and the prevalence of leptospirosis as the Z coordinate, then a contoured graph could be constructed, like that in Fig. 3-2A. The same data could also be plotted as a family of curves by taking, for each locality where the rat population density was a particular value, the water pH value as an X coordirate and the leptospirosis prevalence number as a [illegible], and this is shown in Fig. 3-2B.

Figure 3-1 Listing of the standard set of South American schistosomiasis data used during MOD mapping studies ("NR" means not reported).

DATA POINTS FOR INFECTION RATE (PERCENT) OF SCHISTOSOMIASIS DUE TO SCHISTOSOMA MANSONI IN MAN, 1938-1956, AS GROUPED BY PROVINCES AND SMALL COUNTRIES. EXTRACTED FROM MALEK IN MAY, 1961, STUDIES IN DISEASE ECOLOGY, P. 305-313. INTERPRETED BY MOD STUDY TEAM.

SOUTH AMERICA NORTH OF LAT. -30 DEG.

REGION REPRESENTED BY DATA POINT	REPORTED DISEASE VALUE	INTERPRETED LOCATION LONG.	INTERPRETED LOCATION LAT.	INTERPRETED DISEASE VALUE
LARA + SURR. PRS., VEN.	NR	-70	+11	0
DISTRITO FEDERAL, VEN.	31.6	-67	+11	32
CARABORO PR., VEN.	9.9	-68	+10	9
ARAGUA PR., VEN.	24.8	-67	+10	24
MIRANDA PR., VEN.	10.3	-66	+10	11
MONAGAS + SURR. PRS., VEN.	NR	-63	+ 9	0
BARINAS + SURR. PRS., VEN.	NR	-69	+ 8	0
GUARICO PR., VEN.	30.	-66	+ 8	31
BOLIVAR PR., VEN.	NR	-63	+ 6	0
BRITISH GUIANA	NR	-59	+ 5	0
DUTCH GUIANA	PRESENT	-55	+ 5	1
COLOMBIA	NR	-73	+ 4	0
AMAZONAS PR., VEN.	NR	-66	+ 4	0
FRENCH GUIANA	NR	-53	+ 4	0
RIO BRANCO PR., BRAZ.	NR	-62	+ 2	0
AMAPA PR., BRAZ.	NR	-54	+ 1	0
ECUADOR	NR	-78	- 2	0
AMAZONAS PR., BRAZ.	NR	-65	- 4	0
PARA PR., BRAZ.	NR	-53	- 5	0
MARANHAO PR., BRAZ.	0.46	-45	- 5	1
CEARA PR., BRAZ.	0.94	-39	- 5	1
RIO GRANDE DO NORTE PR., BRAZ.	2.32	-37	- 5	2
PARAIBA PR., BRAZ.	2.32	-37	- 7	8
PIAUI PR., BRAZ.	0.04	-42	- 8	1
PERNAMBUCO PR., BRAZ.	25.17	-36	- 8	26
ACRE PR., BRAZ.	NR	-70	- 9	0
ALAGOAS PR., BRAZ.	20.48	-37	- 9	21
PERU	NR	-75	-11	0
SERGIPE PR., BRAZ.	30.13	-38	-11	31
RONDONIA PR., BRAZ.	NR	-63	-12	0
BAHIA PR., BRAZ.	16.55	-42	-12	17
GOIAS PR., BRAZ.	0.03	-49	-13	1
BOLIVIA	NR	-65	-17	0
MATTO GROSSO PR., BRAZ.	0.007	-55	-17	1
MINAS GERAIS PR., BRAZ.	4.41	-45	-18	4
ESPIRITO SANTO PR., BRAZ.	1.63	-41	-20	2
NORTHERN PARAGUAY	NR	-60	-21	0
SAO PAULO PR., BRAZ.	PRESENT	-49	-22	1
RIO DE JANEIRO PR., BRAZ.	0.10	-43	-22	1
JUJUY PR., ARG.	NR	-66	-23	0
FORMOSA PR., ARG.	NR	-60	-24	0
PARANA PR., BRAZ.	0.12	-52	-24	1
SALTA PR., ARG.	NR	-65	-25	0
SOUTHERN PARAGUAY	NR	-56	-25	0
CHACO PR., ARG.	NR	-61	-26	0
CATAMARCA PR., ARG.	NR	-67	-27	0
TUCUMAN PR., ARG.	NR	-65	-27	0
SANTA CATARINA PR., BRAZ.	0.00	-51	-27	0
SANTIAGO DEL ESTERO PR., ARG.	NR	-63	-28	0
CORRIENTES + ADJ. PRS., ARG.	NR	-57	-28	0
RIO GRANDE DO SUL PR., BRAZ.	NR	-53	-29	0
LA RIOJA PR., ARG.	NR	-67	-30	0

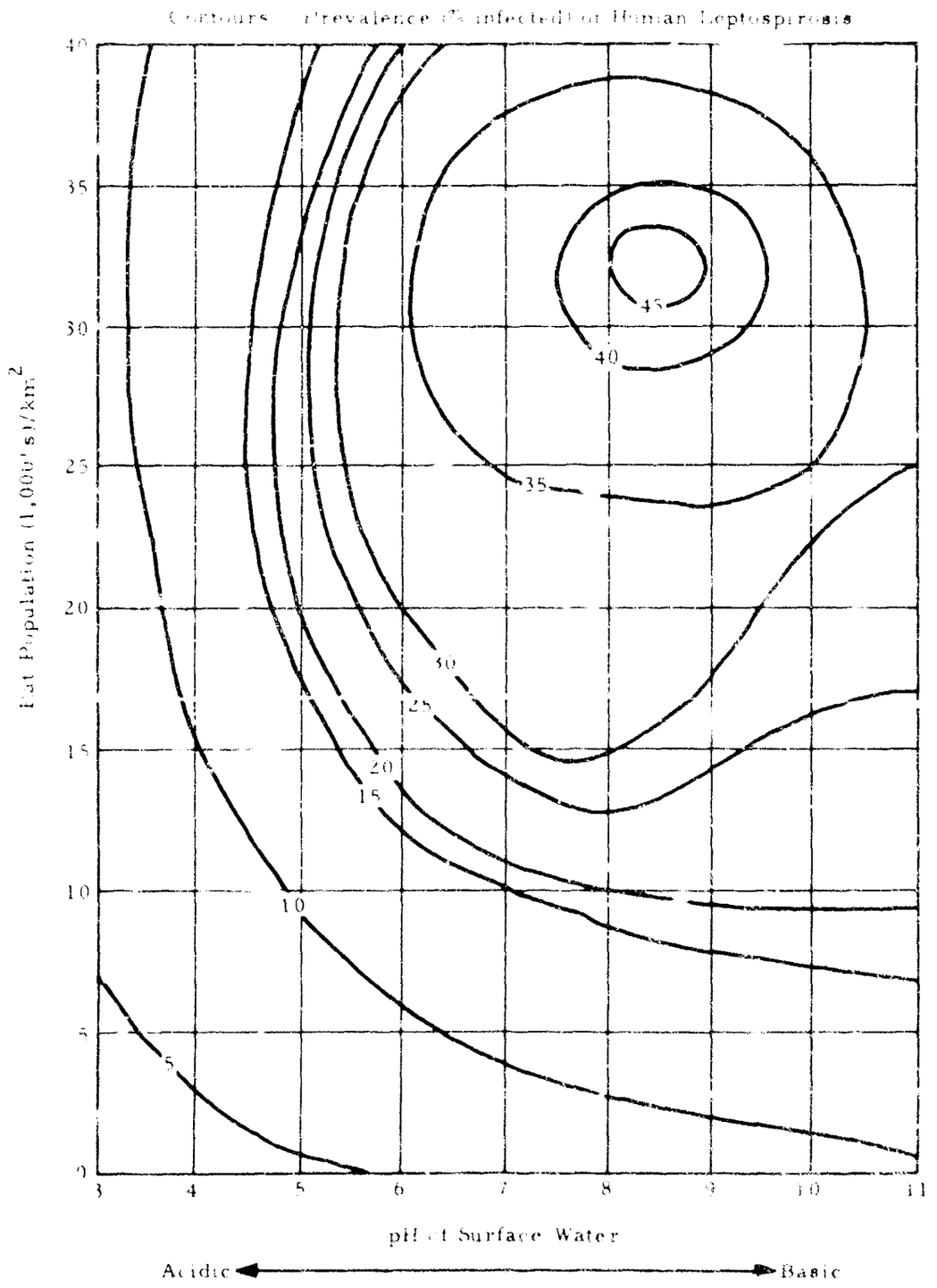


Figure 3-2 Two- and three-variable graphs (A,B,C, and D) illustrating possible disease-environmental relationships: A, a contoured three-variable surface.

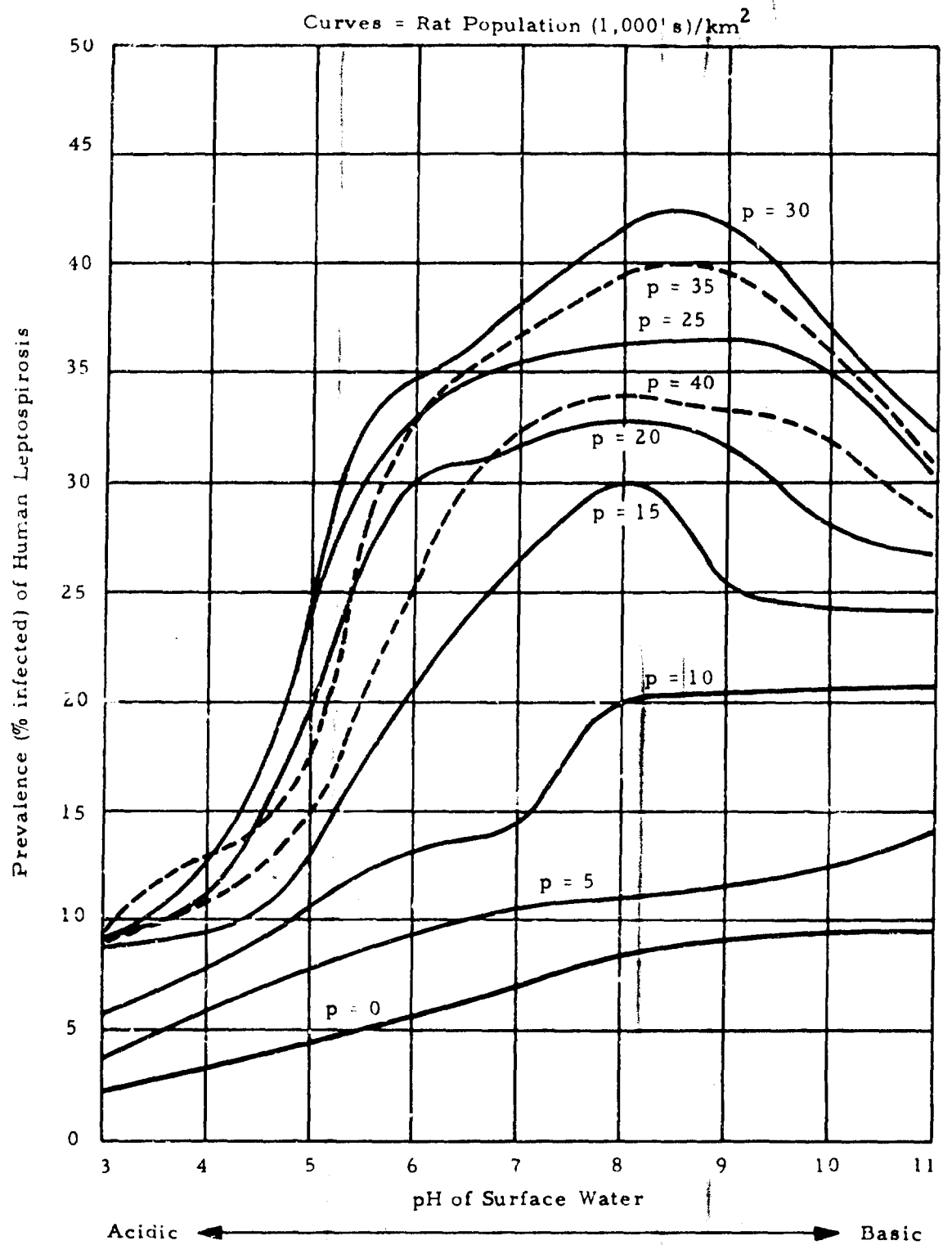
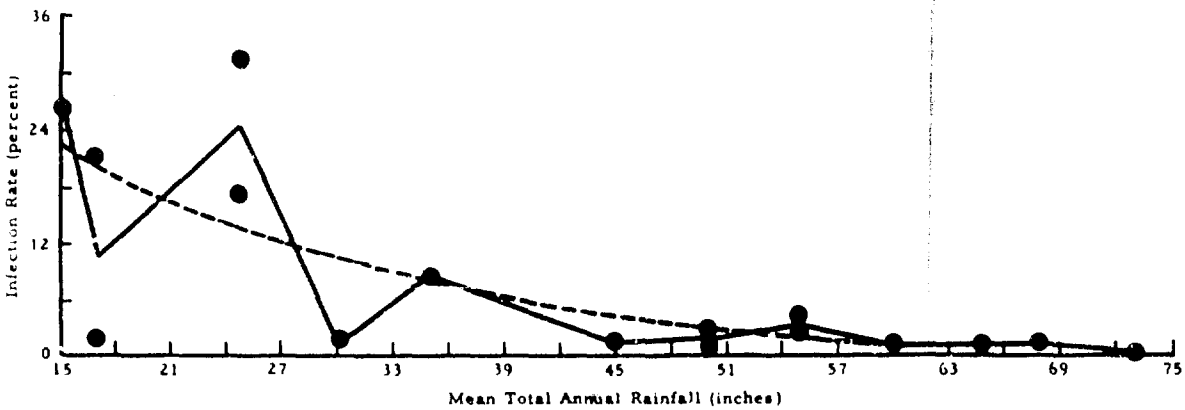


Figure 3-2-B **A family of two-variable curves from a hypothetical leptospirosis situation.**



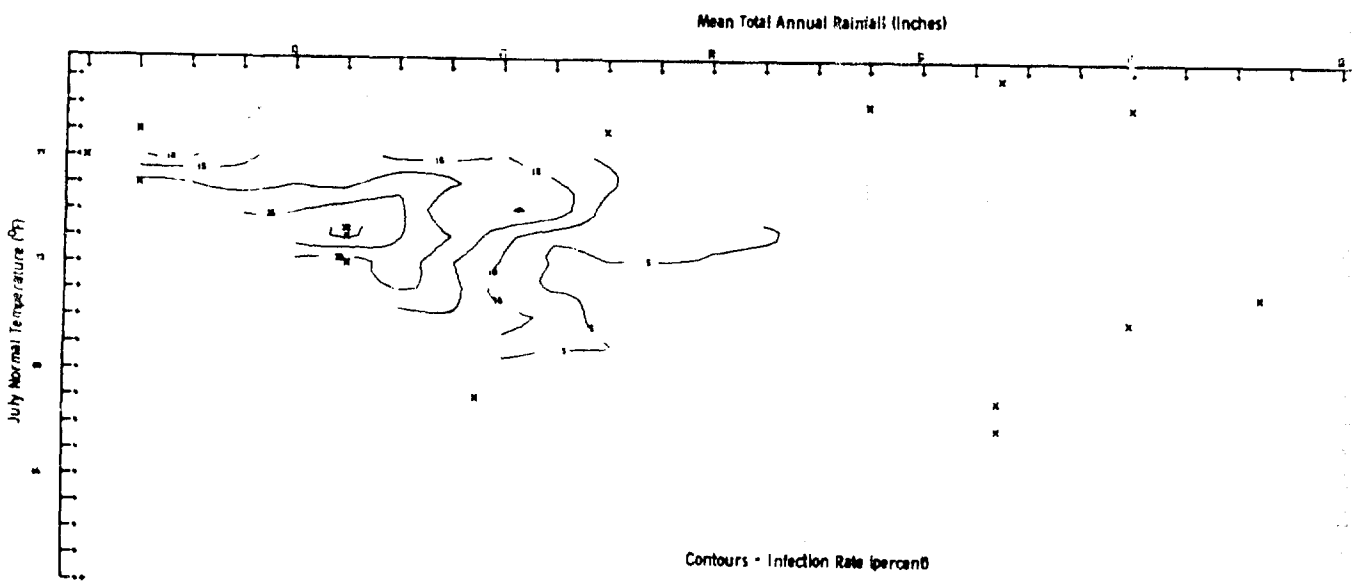


Figure 3-2-C (upper) A two-variable graph, and *3-2-D (lower)*, a three-variable (contoured) graph, both based upon real data: schistosomiasis in Brazil. *(Temperature and rainfall data derived from material in Goode's World Atlas, 12th ed. 1964, copyright by Rand McNally & Co. R.L. 68 S 86; used with permission.)*

Because of the potential power of these graphing methods we also experimented with some actual data. The standard set of schistosomiasis data provided disease factors and two other sets of factors were obtained from existing, readily available maps: annual rainfall (Rand McNally, 1964, p. 97, upper left), and July normal temperature (Rand McNally, 1964, p. 11, lower). Values were extracted from the rainfall and temperature maps at the same longitude and latitude points as each disease data point (i.e., at the center of each respective province). Figure 3-2C shows a graph of two factors, rainfall and infection rate. Two possible curves have been fitted to the data on this graph, one, an exact fit, and the other, a smoothed fit. Figure 3-2D shows a graph of *three* factors: temperature, rainfall, and infection rate. This last graph was contoured to show the possible application of this graphing technique. We emphasize that our purpose here was to explore potentially useful methods; data limitations do not allow conclusions regarding specific disease situations.

The MOD study team did not concentrate on computerized output of this kind because of time and economic constraints, but our limited studies show the technical feasibility and potential usefulness of this kind of output.

3.1.2 GRAPHS

In many fields, where large amounts of data are available, graphs* showing the relationships between two variables are more useful in understanding these relationships than are tables or lists of numbers. Graphs are useful in several ways: when the equation is known, the graph can be used to explain the relationship; when the equation is unknown, the graph can indicate what the equation could be.

---

* A graph is simply a pictorial representation of the relationship between variables and is, in a sense, a substitute for an equation representing this relationship.

Ordinarily, the most useful graph (because of its relative simplicity) is one which represents two variables. Three variables are much more difficult to handle, but can be graphed by means of a "family of curves", each curve representing a two-variable graph for the total situation -- with the third variable held at a particular (constant) value. A more informative representation of three variables, however, is a contour plot, similar to a contour-type map. Study of more than three variables can be facilitated by carefully organized arrays, composed of either two- or three-variable graphs, arranged side-by-side or, perhaps, overlaid one on another.

Any of these graphing techniques has potential (great) value in the study of specific disease-environmental situations. To illustrate their use, consider a hypothetical (realistic) situation in which we have a set of geographic localities where we know the prevalence of human leptospirosis, the density of the rat population, and the average pH of the surface water (ponds, streams, etc.). Suppose, further, that the interrelationship among these factors is such that, broadly speaking, there is more human leptospirosis where two conditions exist together: the surface water is slightly alkaline and there are many rats. The major difficulty in producing graphs to show this relationship comes when one attempts to assign unique single values to the graph points. Perhaps this can be resolved by averaging several points or by narrowly restricting the geographical area from which the graph points are taken.

3.1.3 MAPS

Speaking generally, a map is a representation of spatial or areal relationships on the earth's surface. A map is drawn according to a rigorous, logical, consistent grid pattern and scale, so that there is no non-systematic distortion of size, shape, distance, and neighbors. These are characteristics which other diagrammatic, pictorial, graphic representations do not possess. (For example a cartogram allows non-systematic distortions in the size, shape, and neighbors of regions.) A map has two independent

variables, X and Y, represented by longitude ("LO") and latitude ("LA"), respectively (or latitude/longitude equivalents), and portrays the variations of a third dependent variable, Z, i.e., value ("VAL"), as height above or below a standard plane. Maps and graphs are distinguished in that graphs can use X and Y variables that do not represent (geographic) position. Furthermore, a map is drawn orthogonally to a datum plane; i.e., everywhere on the map, the viewer looks vertically downward toward the center of the earth.

Maps are especially useful when a substantially large volume of (appropriate) consistent, related data is available. Maps are also particularly useful when data have a location characteristic (already represented by two variables, longitude and latitude) and a third variable which consists of the value of the data at that specific location. Under these conditions, a simple two-variable graph is no longer adequate to show the relationships. Maps have proved to be immensely useful tools in all geographically-oriented fields of study. It is in recognition of these important advantages that we have concentrated on maps in our considerations of MOD output.

3.1.4 BLOCK DIAGRAMS

A potentially useful map-like representation of geographically dis-
d data is the block diagram. This differs from a map (as defined cartographically, Lobek, 1958) only in that it is constructed <u>obliquely</u> (rather than perpendicularly) to a datum plane. Because block diagrams resemble maps in so many ways, we will defer further discussion of them until we have considered maps in detail.

3.2 MAP CONSIDERATIONS

Maps are very familiar means of presenting data, especially in simple form, however they can be very complex. In order to establish a basis for

understanding the potential uses and effectiveness of maps -- and their limitations -- the following discussion considers various types of maps, what kinds of data they portray, how they portray these data, and how the maps are constructed.

3.2.1 CATEGORIES OF MAPS

There are so many kinds and uses of maps that it would be virtually impossible to consider them all. The following list of categories (including some items which are not strictly maps) reflects a classification which we developed to catalogue those maps (conventional and computer produced) that might prove useful to the MOD system:

(1) Map index/list/catalog
(2) Outline map
(3) Geographic reference map
(4) Air navigation chart
(5) Photo (mosaic) map
(6) Vertical aerial photograph
(7) Topographic map
(8) Hydrographic (nautical) chart
(9) Base (plat/survey/cadastral) map
(10) Oil company road map
(11) Earth-science map (geologic, soils, glacial and glaciers, weather, climate, mines/minerals, palinspastic, etc.)
(12) Other physical/chemical-environmental map
(13) Biogeographic map
(14) Other biological-environmental map (possibly agriculture, forestry, fishery, etc.)
(15) Disease map
(16) Economic map (agriculture, forestry, fishery, manufacturing, processing, engineering, public works, transportation, communications, trade, commerce, finance, etc.)

(17) Historical map (political/social-historical, paleogeographic, military/naval-historical, etc.)

(18) Other human-environmental map (population, language, religion, etc.)

(19) Extra terrestial map (star chart, etc.)

(20) Other types of maps

3.2.2 USEFULNESS OF MAPS

Why is a map important? Of what use is a map to a scientist studying geographically-distributed factors? Brief consideration has already been given to these questions in the Preface, but primarily in relation to the MOD system. The following statements are a more general response to the questions:

(1) Maps summarize a great deal of information as compared to tables or narrative prose.

(2) Maps enable the scientist to overcome his physical limitations (especially size) and to see the broader spatial relationships and characteristics in the world about him.

(3) Maps are valuable means of communication because the data are presented vividly and in an easily understood visual form.

(4) The use of various graphic methods to represent data permits the pattern of environmental variables to be readily seen.

(5) Maps serve as a powerful means of generalization, aiding in the analysis of spatially/areally distributed data.

- An appropriate general statement (Bick and Johnson, 1967, p. 1) is: "Maps are indispensable in earth science studies because knowledge of the geographic (areal) distribution of quantities, temperature and air pressure, for example, is vital to understanding the processes active on the earth. The use of maps involves competence on the part of both the compiler and the reader with respect to three fundamental factors: an understanding of map scales, how to determine position, and how to present the data in a form that can be readily assimilated."

Underscoring the great potential of maps, the statement has been made that cartography (i.e., the use of maps) has been as important to the development of geographical science as mathematics has been to the development of the physical and engineering sciences (Bunge, 1962, p. 33).

Now that we have emphasized the advantages of maps, let us consider the disadvantages. Maps attempt to present an entire picture from a limited amount of data. All maps are constructed by interpolation techniques, of one sort or another, from a finite (comparatively small) number of observed data points. Map-making techniques are compromises between mathematically rigorous portrayal and psychologically realistic portrayal. Furthermore, every map is, to a greater or lesser extent, schematic and employs conventions -- and every map must be interpreted in the light of at least a general knowledge of how it was made and the conventions employed. These facts do not render maps invalid, but they do impose a responsibility on the user to interpret them intelligently. No one type of map can possess all possible virtues; it is a question of using that type of map which is best for a particular purpose, combining a maximum of relative advantages with a minimum of relative disadvantages, i.e., limitations.

In addition to limitations of the sort we have just described, there may be errors in construction: in locating the X, Y (longitude, latitude) of the data points or in determining the Z (value) of the data points. These can represent observational errors (which depend on the method used to measure the data points' values), sampling errors (when only a limited sample can be taken at unevenly spaced locations), bias error, (in which a person, subconsciously, prefers certain numbers over others), conceptual errors (related to the validity or usefulness of the concept presented on the maps), and, in contour maps, errors due to faulty approximation of the surface by (linear or other) interpolation between known values.

3.2.3 CONVENTIONAL MAPS

Our purpose here is to point out several important characteristics of

conventional maps, characteristics that are not necessarily related.

A map portrays items symbolically, and much of cartographic symbolism has grown up over the years to the point that it is now a well-established, standardized set of conventions. One of the most important of these cartographic conventions concerns the base information placed on a particular map to show that minimum amount of geographic reference information (coastlines, political boundaries, cities, rivers, etc.)necessary for the viewer of the map to correlate the distribution of the factor being mapped with familiar landmark points on the earth's surface. (Maps which are prepared by a computer are usually either drawn upon or laid over a base map which already contains most of these data.)

The kinds of data which can be mapped are unlimited, so long as the data can be phrased in terms of X, Y, Z triplets (longitude, latitude, and value). The major problem comes in selecting data points so that the resulting map surface is most informative and of "reasonable" appearance. For example, no fundamentally different techniques of mapping are involved when making a contour map of (land based) diseases in a coastal region than are involved when making such a map for an interior region. But in the case of the coastal region, the cartographer (human or computer) influences the map he is making by modifying his set of data (*not* his cartographic methods) to include a large number of data points out in the ocean, each point with a value equivalent to "no disease present".

Logically, a map represents only one dependent variable (or one disease-environmental factor), but more than one variable can be represented either by a series of maps (each displaying a distinct or unique statistical surface) or by overprinting the mapped patterns of several such variables onto the same base sheet.

3.2.4 DISEASE MAPS

A disease map is one that shows the geographic (areal or spatial) distribution of some clearly defined aspect (qualitative or quantitative) of the total disease situation. Disease maps are abstract statistical surfaces, constructed artificially according to the same conventions that define the construction of conventional maps. The disease maps which have been produced to date show principally the distribution or occurrence of a particular item in the total disease situation, often with extensive narrative comments on the maps themselves and/or in the legends (and the use of such verbage indicates a failure to present adequately, in map form, the medical information).

Disease maps are closely related to maps of other abstract statistical surfaces, e.g., population density, rate of change of population density, etc., which cannot actually be seen or observed directly in the field, but which must be calculated from field observational data -- in contrast to road maps, topographic maps, type-of-bedrock geologic maps, etc.

The use of disease-environmental maps as a research tool is based on the assumption that coincidence in the distribution patterns of two mapped factors indicates a relationship (causal, associative, or coincidental) between those factors. One of the most clear-cut instances in which mapped patterns have been used to indicate important disease-environmental relationships is to be found in the conclusion (Burkitt, 1962, p. 77-78) that the distribution of Burkitt's tumor, in Africa, occurs where, simultaneously: the altitude is less than 5,000 feet; the seasonal mean temperature is always greater than 60°F; and the total annual rainfall is greater than 20 inches (Fig. 3-3). From the apparent interrelationship among these factors, Burkitt has suggested that this tumor may be caused by an insect-borne viral agent. Another clear-cut instance in which a disease distribution pattern (goiter) matches that of an environmental factor (iodine content of drinking water) is shown in Figure 3-4. A third illustration (Kratchman and

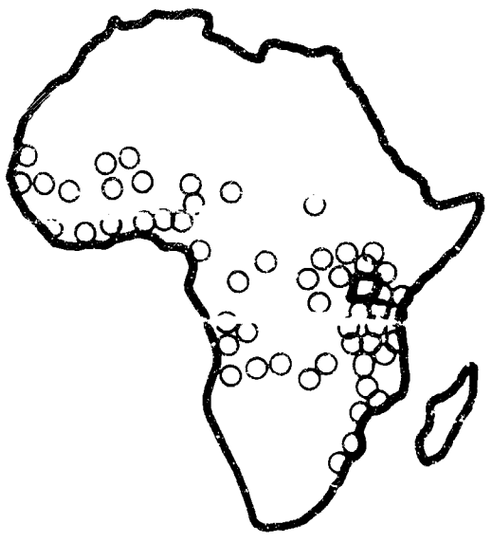


Figure 3-3-A Distribution (circles) showing location of Burkitt's tumor in Africa.

from Postgraduate Medical Journal, Vol. 38, p. 71; reproduced with permission of the Editor and of the Author, D. Burkitt.

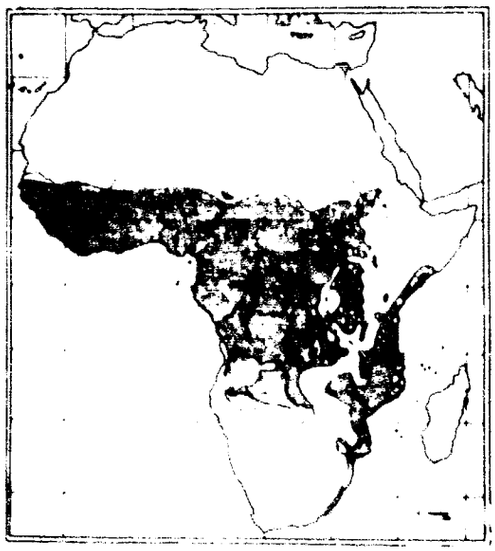


Figure 3-3-B Area (shaded) where the following three conditions are met simultaneously: altitude is under 5000 feet, seasonal mean temperature always exceeds 60°F, and total annual rainfall exceeds 20 inches.

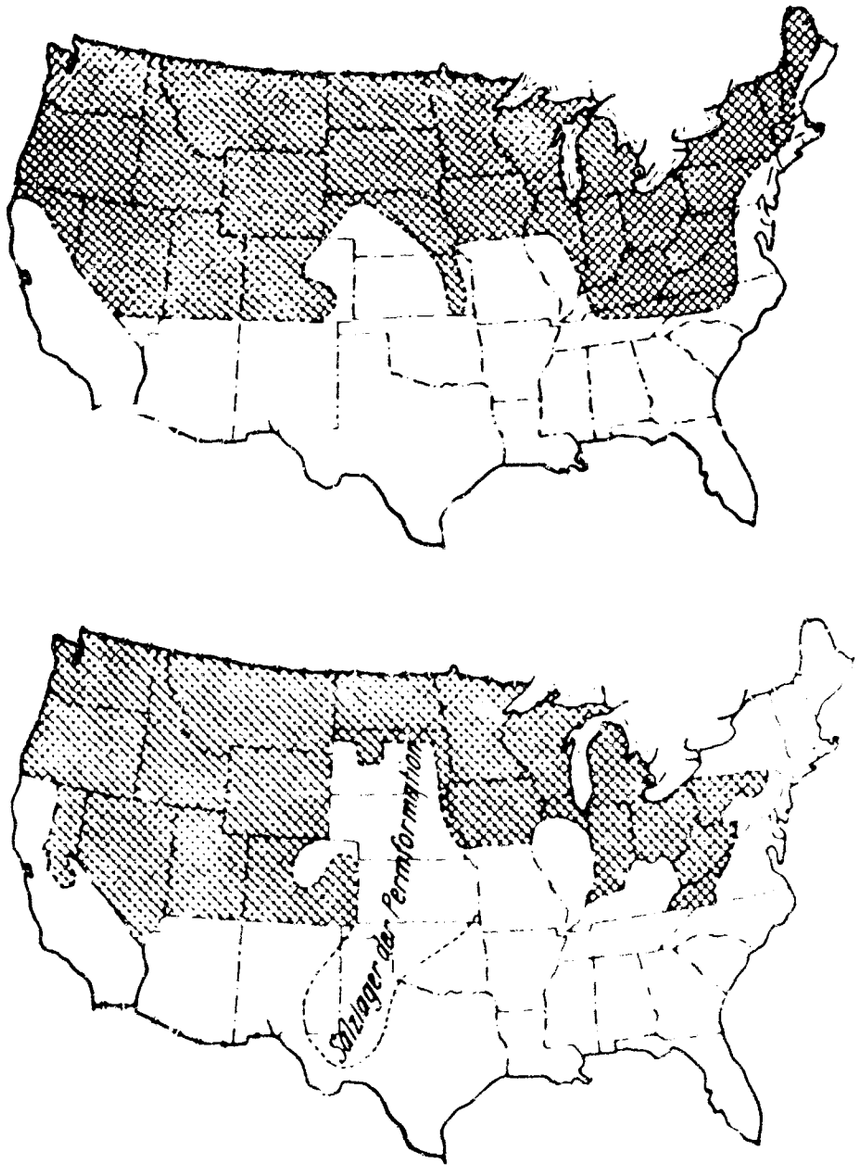


Figure 3-4 Relationships between goiter and iodine content of drinking water: *A (upper)*, area (shaded) of United States, about 1920, with goiter "frequent" (five or more cases per thousand); *B (lower)*, area (shaded) with iodine content of drinking water "low" (less than 0.23 per liter).

from History of Geography of Diseases by Henschen, F., 1962 (English translation by Tate, 1966), a Seymour Lawrence Book published by Delacorte Press; used with permission.

Grahn, 1959) involves the correlation of deaths from congenital malformations with higher-than-average levels of environmental radiation. A more detailed consideration of the use of disease maps, particularly the types that could be produced by the MOD system, is given in Section 8, Output Usage.

3.2.5 SYMBOLIC REPRESENTATION ON MAPS

Data can be represented graphically on a map in three basically different ways, and these may be combined. One may use:

- Dot-type symbols, such as actual dots or points, numbers, letters -- even small pictures. These can be considered as zero-dimensional symbols since, in practice, they approximate a geometric point. (The term dot-type is used here in its literal, descriptive sense. We realize that, to a cartographer, a dot-type map is a particular kind of statistical map which shows density distributions by dots. We are using "dot-type" as synonymous with "point value-" or "data point-" or "point-type".)
- Shading-type symbols, such as various intensities of grey, or various colors, or patterns. These can be considered as two-dimensional symbols since, in practice, they approximate a geometric planar area. (Some maps using shading-type symbols are also known as choropleth maps, others as dasymetric maps.)
- Contour-type symbols, or contour lines (also known as isarithms, isolines, or isopleths). These can be considered as three-dimensional symbols since a set of contour lines, in practice, approximates a geometric curved surface.
- A fourth method of representing information on a map utilizes flow-line symbols such as directional arrows (Fig. 3-5), which may be considered as one-dimensional symbols, approximating geometric lines. This method is mentioned only in passing as it has but limited application; it must be used in combination with one of the other three types of symbols to be meaningful.

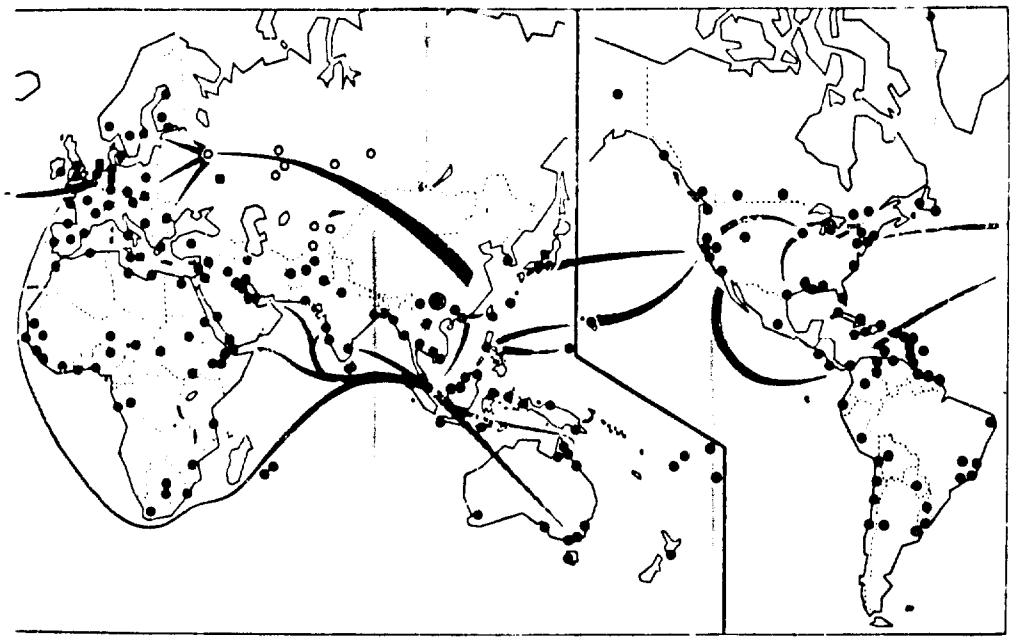


Figure 3-5 Flow line-type maps illustrating the routes by which "Asian flu" spread over the world during February 1957 - January 1958: star (in southern China) indicates probable origin; black (and white) dots show the first wave of cases.

from History of Geography of Diseases by Henschen, F., 1962 (English translation by Tate, 1966), A Seymour Lawrence Book published by Delacorte Press; used with permission.

It is appropriate to reconsider the meaning of data in the light of the three kinds of symbolic representations we have just presented. When numerical data are presented on maps, cartographers commonly speak of these as *statistical maps*. Many of the data which relate to disease-environmental situations are numerical (quantitative), many are qualitative. The following division of symbols, on the basis of qualitative versus quantitative, may be helpful.

SYMBOL	QUALITATIVE	QUANTITATIVE
Point and Line *(dot-type)*	roads towns	dot distribution density of population flow lines
Area *(shading-type)*	vegetation type land-use type land-form symbols	chorapleth and other shadings where numerical values have been assigned circles shadings
Contour	nominal or ordinal level data	isometric data

3.2.5.1 Dot-type maps Dot-type (data point) maps are useful tools for the study of various environmental factors, especially their qualitative aspects, i.e., whether a particular factor is present (yes) or is not present (no) at a particular place. (Quantitative aspects can be shown, however, when different kinds of dots ((size, color, etc.)) are used to indicate differences in amount.) Probably the best known dot-type maps are those showing biogeographic distribution of various species, including human population distribution maps.

To construct a dot-type map the cartographer specifies exactly what disease or environmental factor he intends to map, and just how he will draw

the map. Next, he obtains an internally consistent, relevant set of data points, each of which are expressable as longitude, latitude, value (LO, LA, VAL) triplets. Then he selects a sheet of paper appropriately gridded for longitude (LO) and latitude (LA), places a dot on the grid according to the (LO, LA) of the point, and writes the point's value (VAL) next to that dot. Finally, he divides the total range of VAL's represented on the entire map into several groups or intervals, selects an appropriate dot-type symbol for each interval, and draws over each dot the appropriate symbol for its VAL.

In this consideration of dot-type (point-value) maps we have used the sort of technique which would be necessary for any computer programmed system. In practice the (human) cartographer might not locate dots by latitude and longitude, particularly if the density of data points was such that this degree of precision was meaningless. The density of dots might be determined by the data being plotted, e.g., one dot = one hundred people, in which case the dots would be placed to represent the pattern in reality. If there were only 100 people in a county, the single dot would be placed where most people lived. The center of the (geographic) unit would be chosen only if the distribution were uniform.

Figures 3-6 through 3-10 illustrate various dot-type maps -- some from published works, other produced by the MOD group. Figures 3-6 and 3-7, from published papers, are dot-type maps that portray the distribution of several environmental factors. Figure 3-8, in its published form used dot-type symbols of different *colors* as well as shapes to show the distribution of leptospiral serotypes. Figure 3-9 presents our standard set of (schistosomiasis) data as two manually drawn, dot-type maps; Figure 3-10 as computer/line-printer output (using the Kansas Geological Survey trend-surface program).

3.2.5.2 Shading-type Maps The "statistical surface" represented by a shading-type map consists of a series of essentially horizontal planes that have different elevations and that are separated by vertical cliffs = escarpments (i.e., a step function). The different elevations are represented by variations in shades of grey, or patterns, or color.

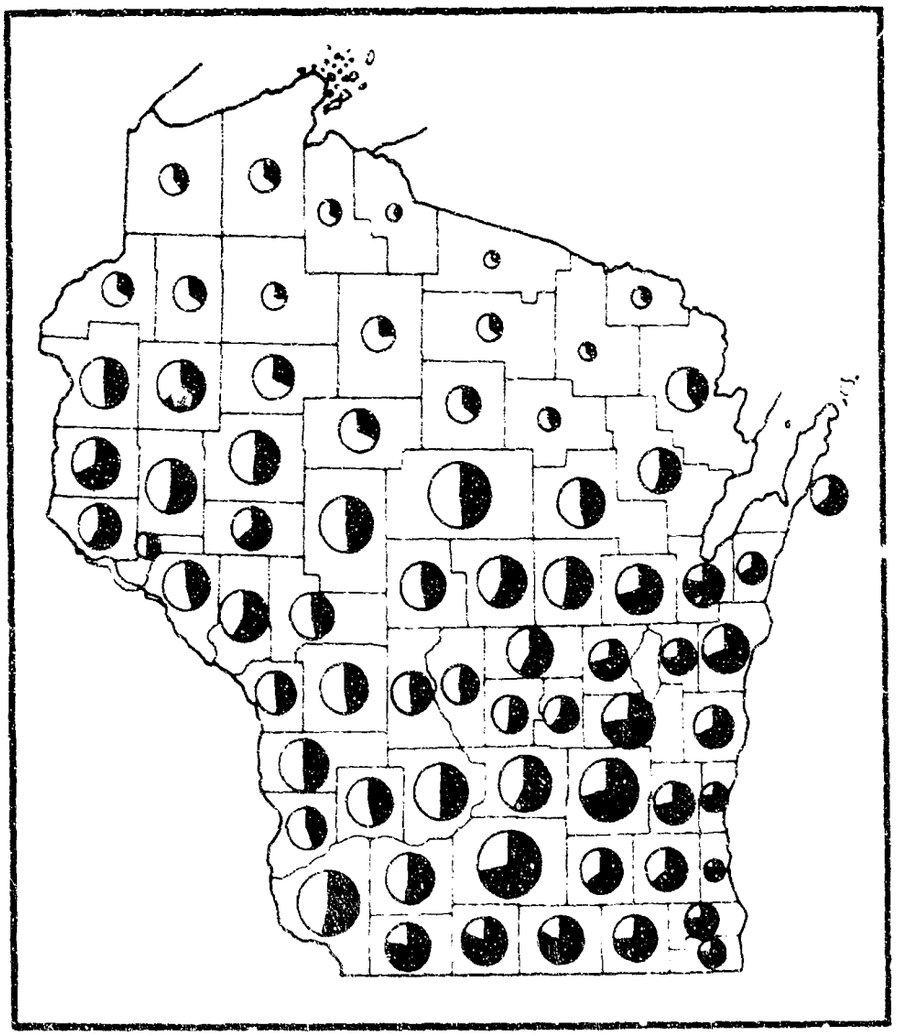
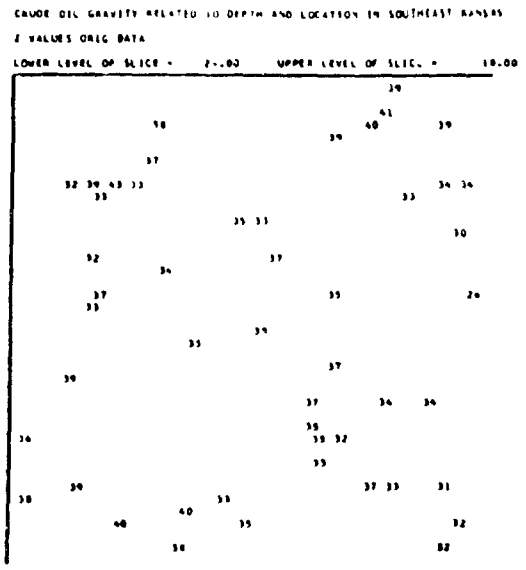


Figure 3-6 Published, manually drawn, dot-type map showing the amount of land in farms (related to the size of the circles), also the percentage of that land available for crops (related to the area of the black sectors within the circles).

from Elements of Cartography, 2nd ed., by Robinson, A. H., 1960, published by John Wiley and Sons, Inc., New York and reproduced with permission.



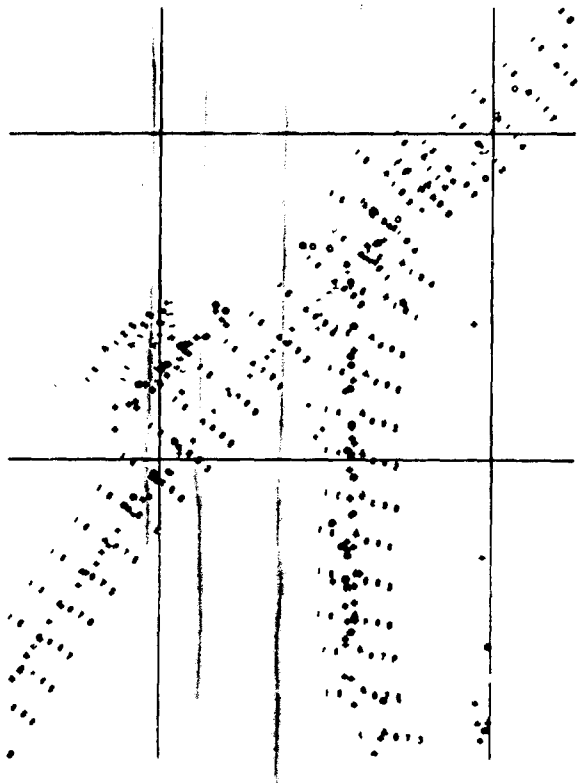


Figure 3-7 Examples of machine-drawn dot-type maps: *A (upper)*, showing crude oil gravity values, plotted on a line-printer (Harbaugh, 1964, p. 57; courtesy of State Geological Survey of Kansas); *B (lower)*, showing aircraft positions during an interception, plotted on a Benson-Lehner plotter (courtesy of Bell Telephone Laboratories).

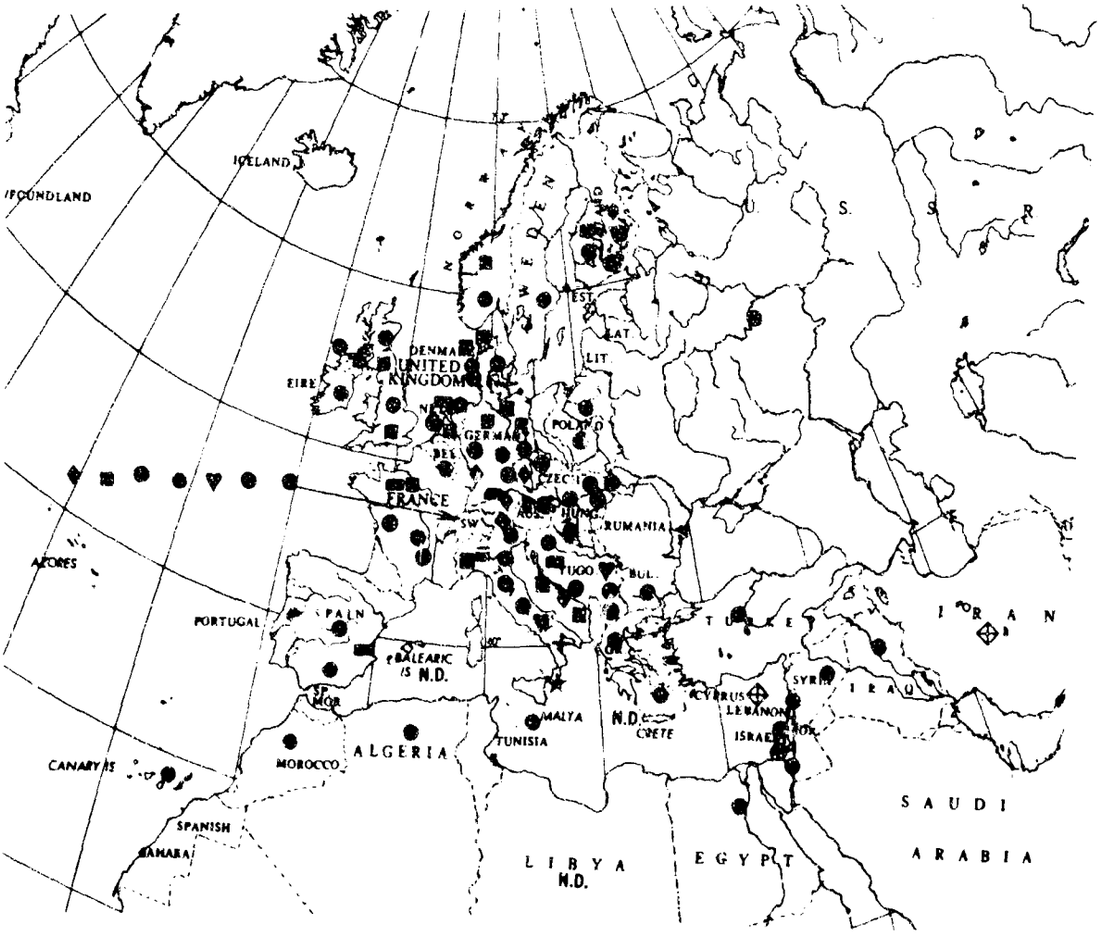
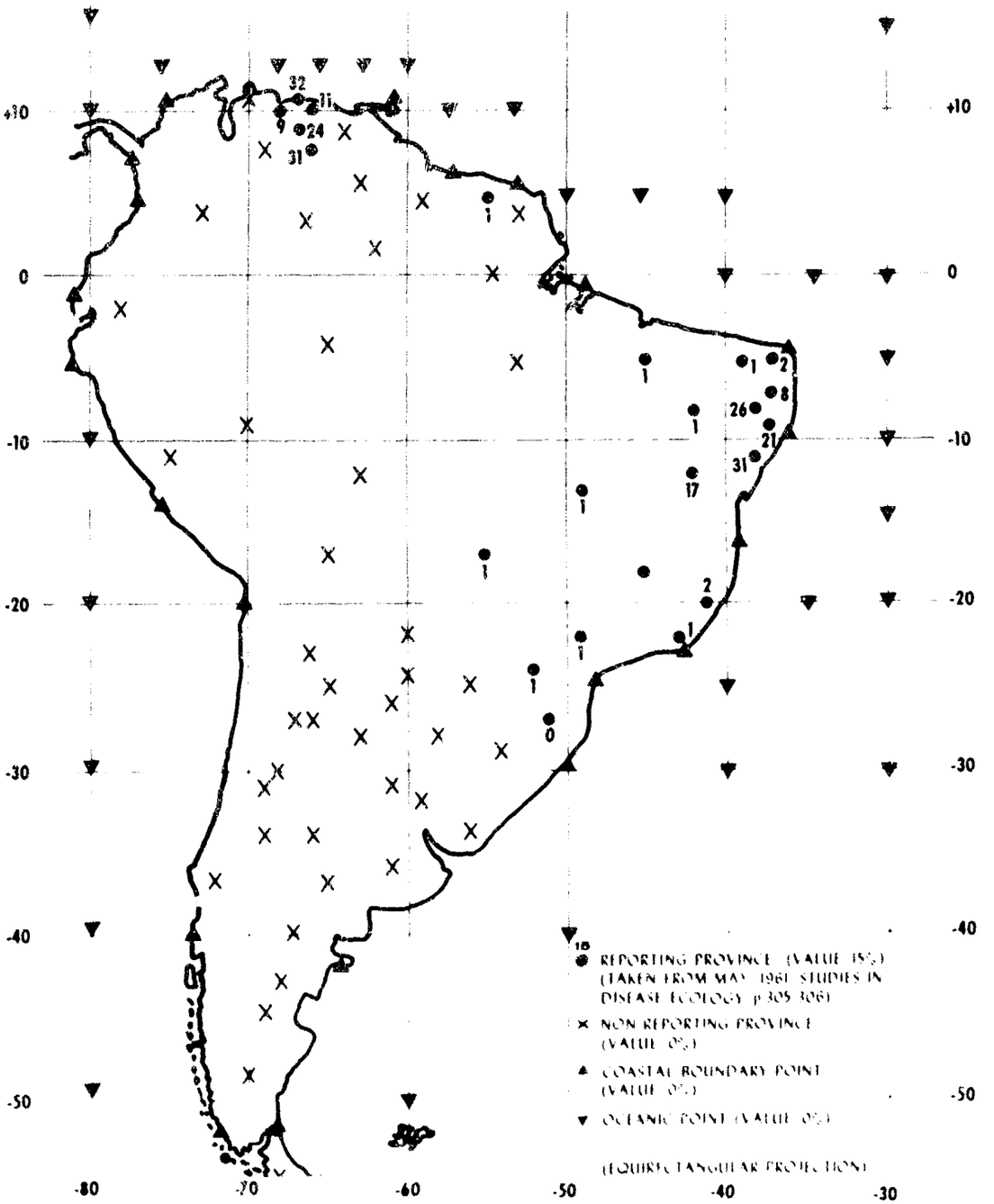


Figure 3-8 Section of published, manually drawn, dot-type map portraying the occurrence of various leptospiral serotypes in different countries; the presence of each serotype is symbolized by a dot of a different shape and color on the original map.



Figure 3-9 Dot-type maps (*A and B*) drawn manually from standard set of South American schistosomiasis data (Figure 3-1) as part of MOD study; *A*, simulating MOD machine-produced disease map.



Figur 3-9-B Same data shown in *Figure 3-9-A*, but following a more conventional cartographic format.

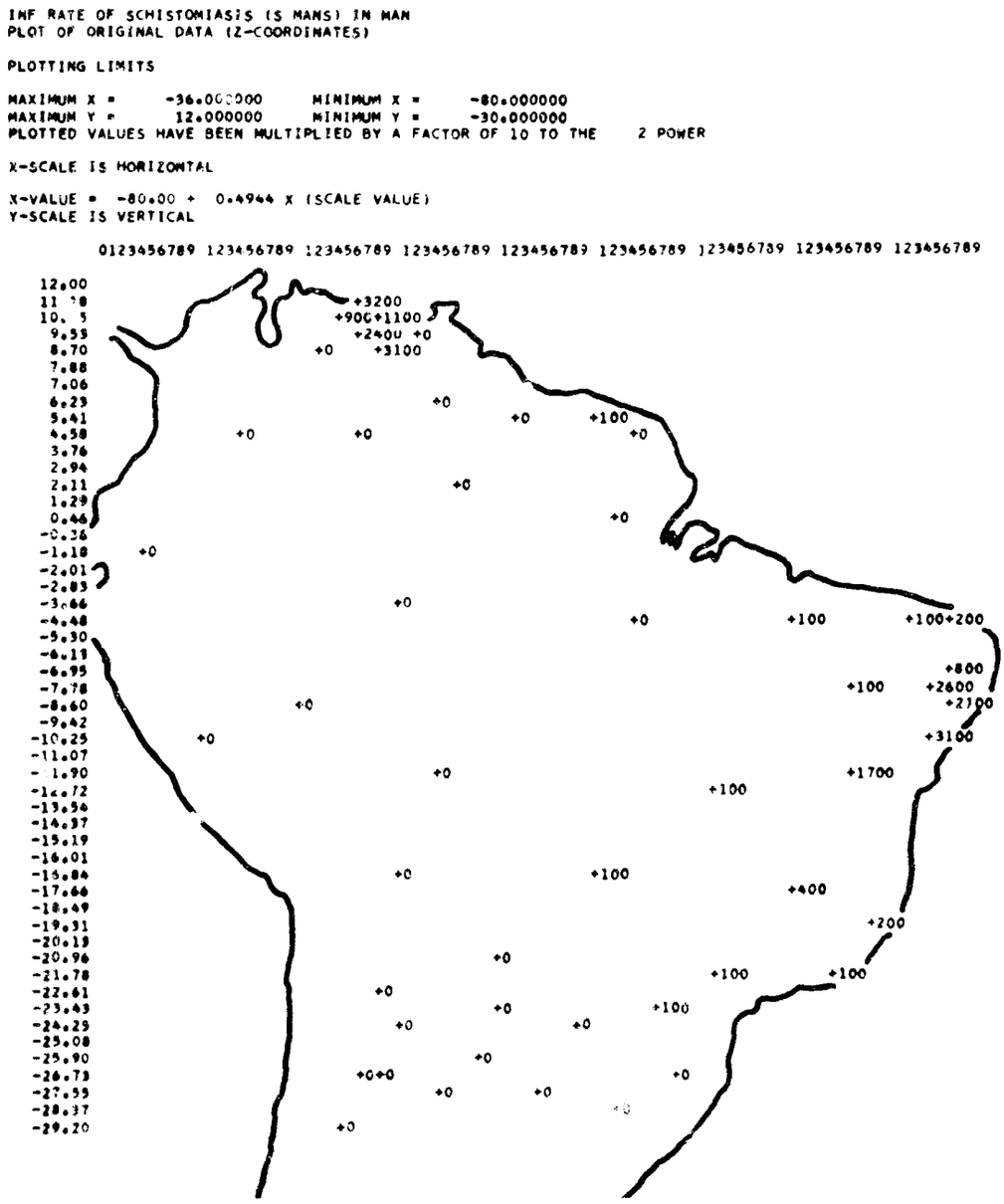


Figure 3-10 Dot-type map produced during MOD study from the standard set of South American schistosomiasis data (Figure 3-1), utilizing Kansas Geological Survey trend-surface program running on an IBM 7090 computer with output on a line printer (outline of the continent was added manually).

Shading-type maps have also proved to be quite useful tools in studying various environmental factors, especially in relating a qualitative factor to an area location. (As with dot-type maps, quantitative aspects can also be shown.) Well-known examples of shading-type maps are the large- to small-scale bedrock geology maps published for most states and countries by governmental geological surveys.

To construct a shading-type map the cartographer specifies exactly what disease or environmental factor he intends to map, and just how he will draw the map. Next, he obtains an internally consistent, relevant set of data points, each of which must be expressable in the form of (LO, LA, VAL) triplets. Then he selects a sheet of paper appropriately gridded for LO and LA, draws the outlines of the units (political or otherwise) on the (LO, LA) grid, and enters the VAL for the area (determined by grouping procedures). Finally, he divides the total range of VAL's represented on the entire map into several intervals or groups, selects an appropriate shading-type symbol for each interval, and shades or colors each unit with the appropriate symbol for its VAL.

Figures 3-11 through 3-15 illustrate various shading-type maps. Some of these are from published works, others were produced by the MOD group. Figure 3-13, taken from the published medical literature, shows how shading techniques can be used to represent various medical data. Figure 3-14 shows the standard set of (schistosomiasis) data presented in the form of a shading-type map. Figure 3-15 portrays that same data as output by a computer/line-printer configuration, using a simple program which we prepared.

3.2.5.3 Contour-type Maps Contour-type maps have proved to be very useful tools for the study of various environmental factors, particularly those considered by the several earth sciences. They are also very well suited for the study of many disease situations. This is because contour maps present quantitative as well as qualitative aspects.

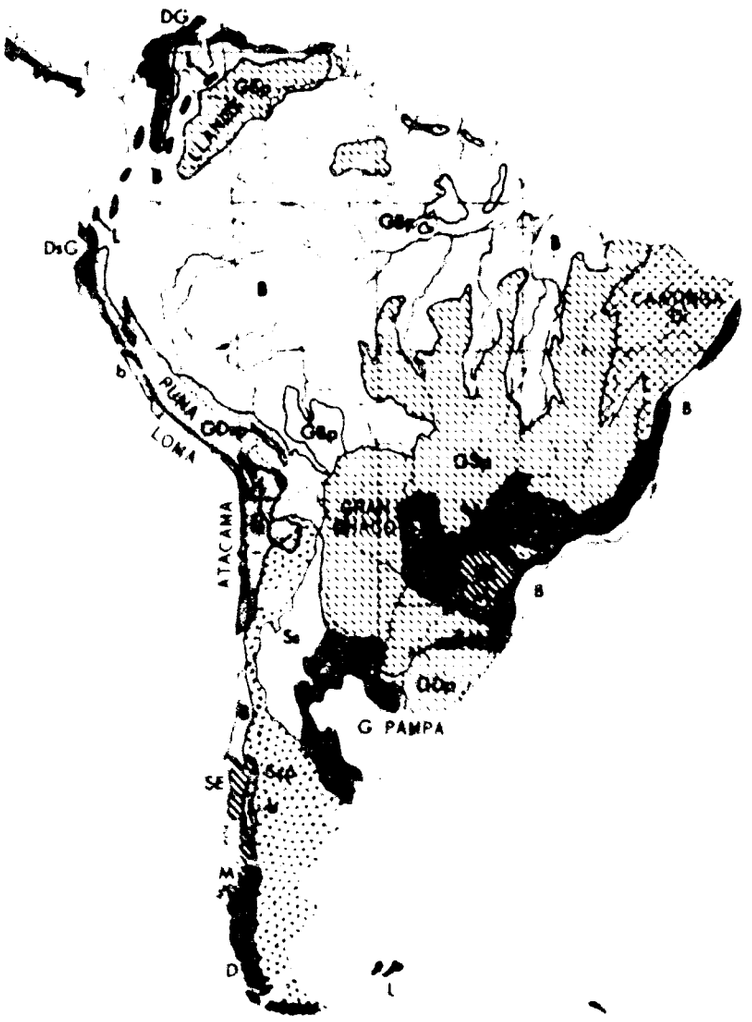
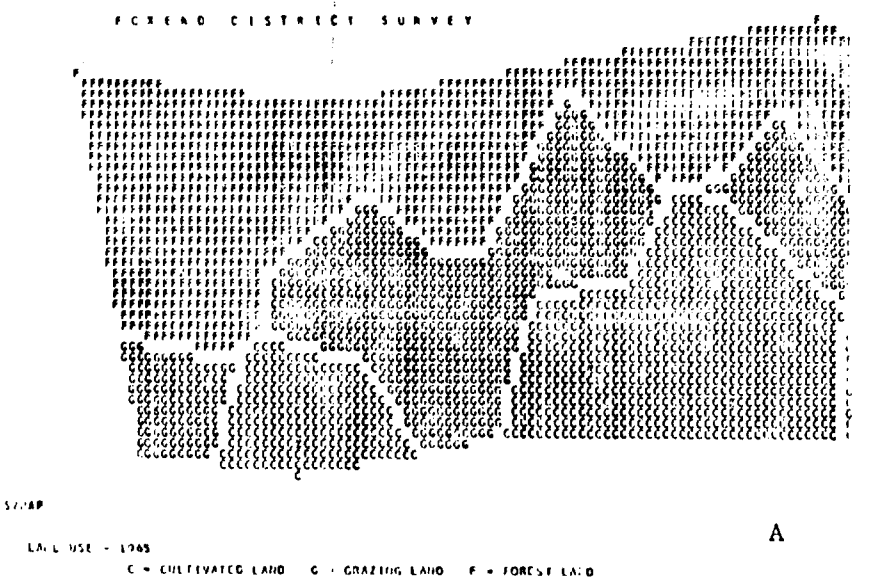


Figure 3-11 Section of published, manually drawn, shading-type map showing types of natural vegetation. On the original map each different shading-color pattern represents a different kind of vegetation.

from Goode's World Atlas, 12th ed., 1964
Copyright by Rand McNally & Co., R.L. 68 S 66; used with permission.



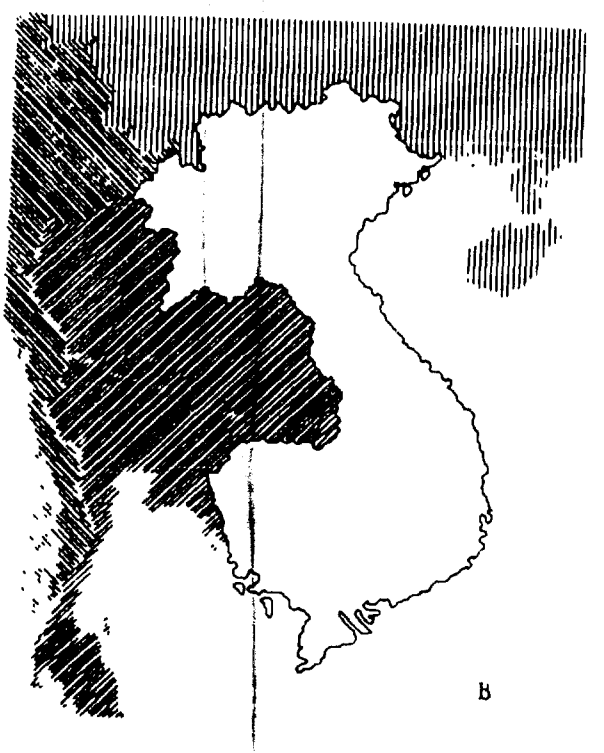


Figure 3-12 Examples of machine-drawn, shading-type maps: *A*, done by computer/line-printer configuration (Fisher et al, 1967, courtesy of Laboratory for Computer Graphics, Harvard University); *B*, done by computer/plotter configuration *from Computer Representation of Planar Regions by their Skeleton, by Pfaltz and Rosenfeld: Communications of the ACM, Vol. 10, No. 2 (Feb.), 1967; reproduced with permission.*

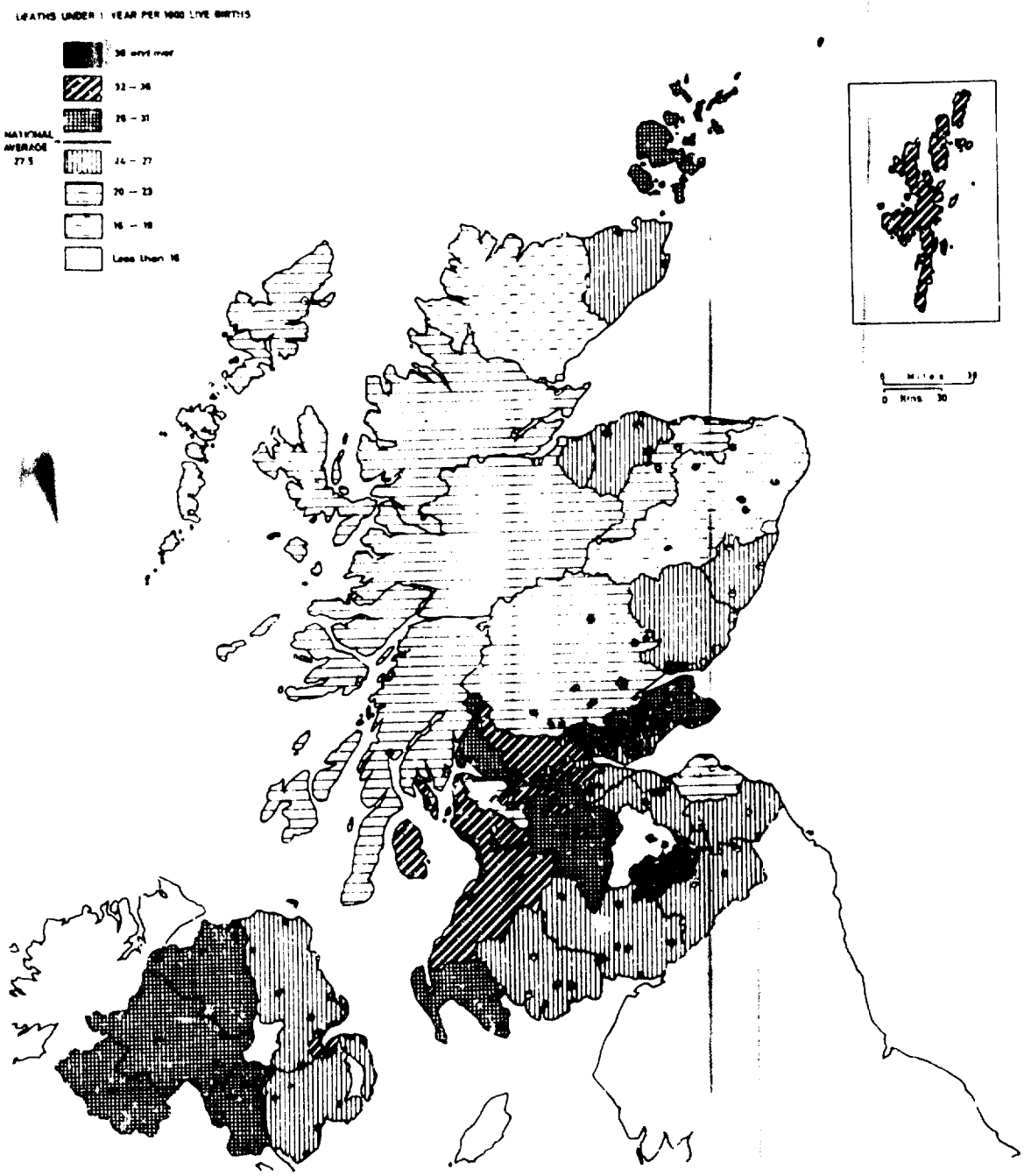


Figure 3-13 Published, manually drawn, shading-type map displaying statistics that represent infant mortality in part of Great Britain.

from National Atlas of Disease Mortality in the United Kingdom, 1963, by Howe, G.M., reproduced with permission of the publisher, Thomas Nelson and Sons Ltd., Middlesex.

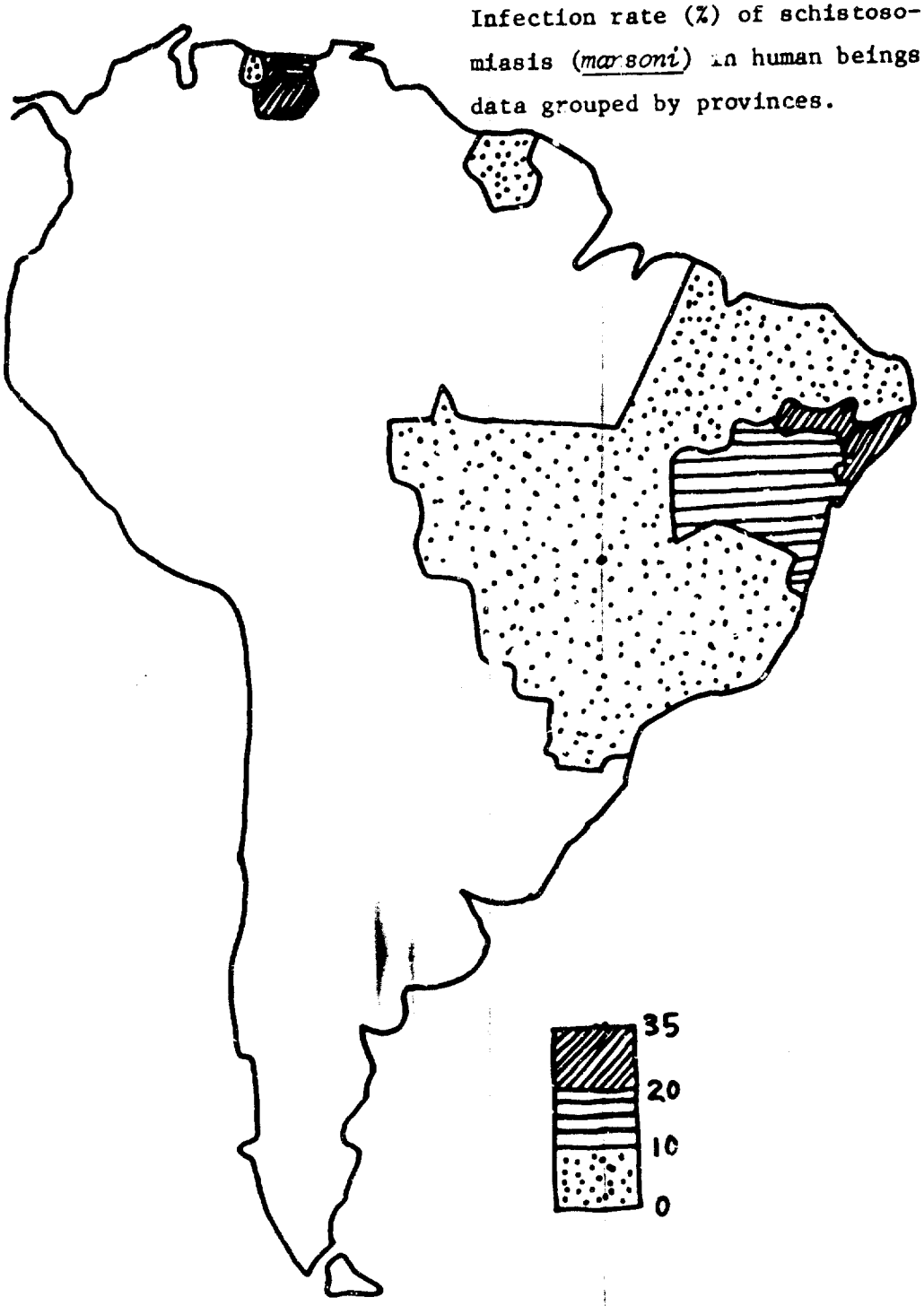


Figure 3-14 Manually drawn, shading-type map showing the same standard set of South American schistosomiasis data (Figure 3-1) as presented in Figure 3-9.

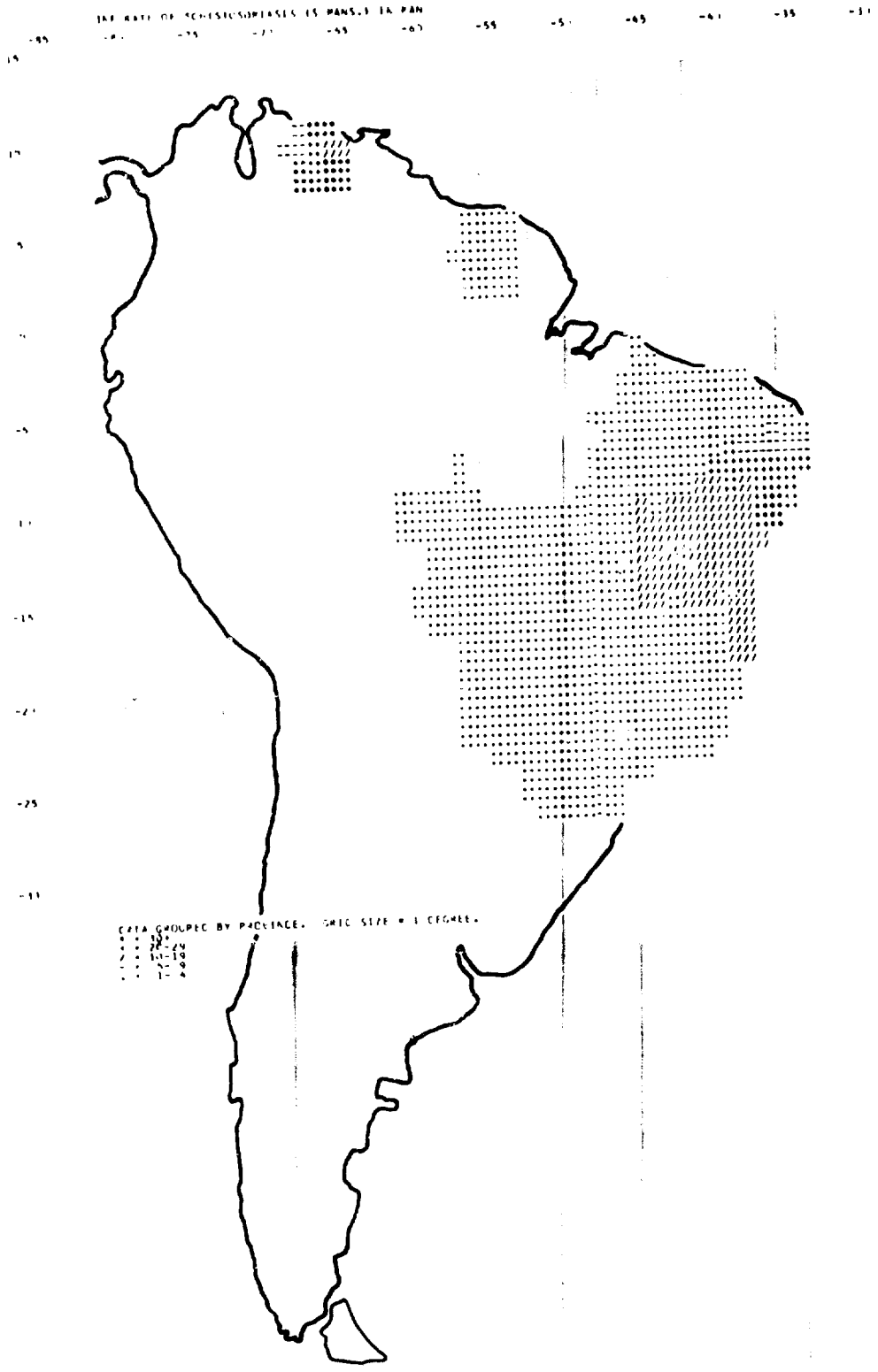


Figure 3-15 Shading-type map of standard set of schistosomiasis data (Figure 3-1) output on a line printer after processing by an IBM 7090 (outline of South America added manually).

3. Output Analysis

Dot-type and shading-type maps are (intuitively) understood by most biomedical and data-processing oriented people with whom the study team has conversed, but this was not the case with contour-type maps. For this reason, and because contour maps are so important in the MOD project, we present here a somewhat detailed explanation of what contour-type maps mean and how they are constructed.

Well-known examples of contour-type mapping include the large-scale topographic maps so widely used in the earth sciences, and weather maps. Basically, a contour-type map is a device by means of which a three-dimensional, complex geometrical figure (real or imagined) can be represented on a two-dimensional, simple plane surface. The map accomplishes this by a "set of lines", i.e., contours (sometimes called isolines, isarithms or isopleths) which outline, according to well-defined rules, the shape of that complex geometrical figure. Each contour is a line drawn on the map or chart connecting points of equal value. This line often represents points of equal elevation above (or below) some assumed base elevation, but it may also represent points of equal temperature or humidity or population density or disease prevalence.

Because the geometrical figure being represented by the map is three-dimensional, that figure can be treated as a set consisting of many ordered triplets of numbers (X_i, Y_i, Z_i). For each specific pair (X_a, Y_a), one, and only one, Z value (Z_a) exists; i.e., $Z = F(X, Y)$.

Contour techniques can be used to represent the form of any geometrical surface of the form $(X_i, Y_i, Z_i = F(X_i, Y_i))$. The variables X and Y, theoretically, can represent values for any conceivable independent disease-environmental factors. Under these conditions the result is a <u>graph</u> showing the relationship among three disease-environmental factors. In order to make a contour <u>map</u> of a particular disease-environmental factor, X is taken to be the LO (longitude) of a geographic point, Y the LA (latitude), and Z the VAL (value) of the specific factor at that particular (X,Y)

point locality. Thus, in making a map, the X, Y, Z triplet becomes a special case; it becomes a LO, LA, VAL triplet.

Now that we have considered these general principles, let us examine a standard topographic contour map. On such a map, LO (=X) and LA (=Y) are obvious and need no elaboration. VAL (=Z) is taken to be the elevation in feet above or below the datum plane of mean sea level. The relationship between an actual land surface and its representation as a contour-type map is illustrated in Figure 3-16. The remainder of this discussion will be based upon actual construction of a contour-type topographic map. Although this illustration uses VAL's of elevation, any disease-environmental factor which could be assigned a unique value (VAL_a) at each specific geographic locality could be contour mapped in precisely the same way, e.g., to show the infection rate of Schistosoma mansoni in man, or the mean total annual rainfall.

To describe the production of a contour-type map, we will use a specific example. We will construct a contour map (with 10-foot contour intervals) showing land elevation (in feet) above mean sea level over a small area containing a twin - peaked hill (Fig. 3-17A). We will use the data contained in Fig. 3-17B, which were actually taken from Fig. 3-17A, and a square (LO, LA) grid. [▶ in the margin mark discrete steps.]

▶ First the cartographer specifies exactly what disease-environmental factor he intends to map and just how he will draw the map (selects contour interval, etc.). Next, he obtains an internally consistent, relevant set of data points expressed in the form of LO, LA, VAL triplets, and selects a sheet of paper, appropriately gridded for LO and LA.

▶ Then, just as with dot-type or shading-type maps, he plots each data point on the (LO, LA) grid and writes the point's VAL next to that dot. By this operation, the data of Figure 3-17B become the dot-type map shown in Figure 3-18. But from this step on, the procedure differs from those described before.

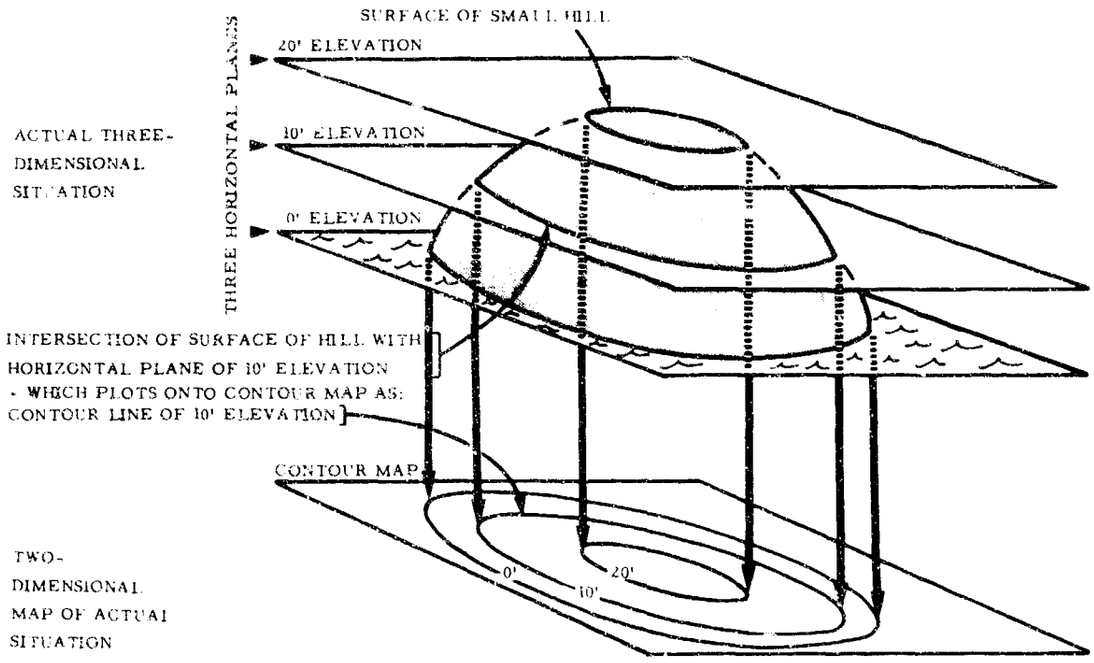
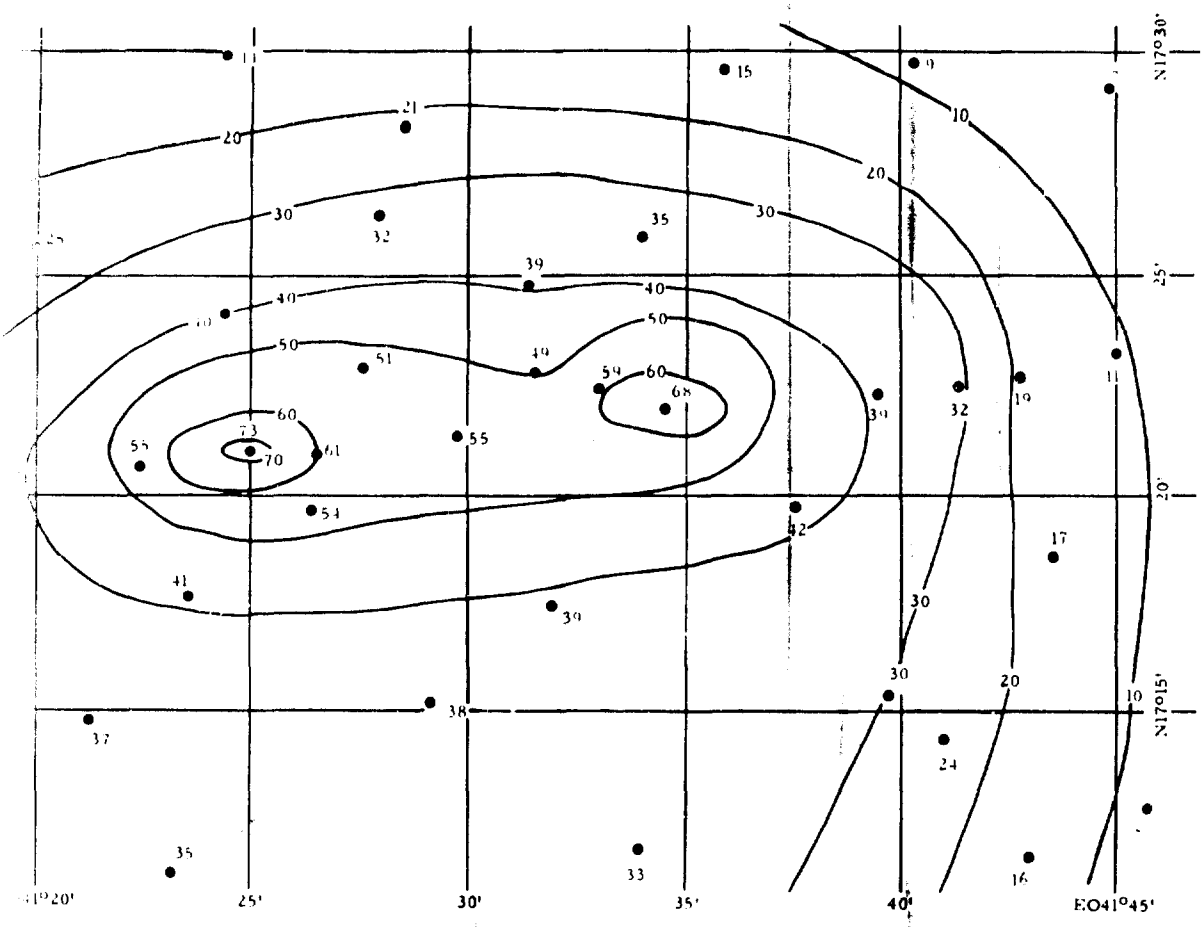


Figure 3-16 Relationship between an actual land surface and its representation as a contour-type map.

After all the data points have been inserted, the cartographer scans (visually) to note the highest- and lowest-valued points, also any broad trends that the data point values may suggest. In our example (see Fig. 3-18), the highest-valued data point is the 73 foot point near $41^{\circ}25'$, $17^{\circ}22'$, and the lowest-valued data point is the 5 foot point near $41^{\circ}45'$, $17^{\circ}29'$. The data points indicate a high area in the left-central and central-central portions of the map and imply that the surface slopes downward sharply north, northeast, and east of this high area, and more gently south from the high area. Another important observation is that the data points appear to be most densely distributed in the upper-left quadrant of the map.

After this preliminary evaluation, the cartographer decides which contour line will probably be easiest to draw and, at the same time, will suggest the overall shape of the surface being contoured. The highest-valued data point is commonly chosen and successively lower-valued contours drawn around it. (Occasionally the lowest-valued data point is a better choice.) Sometimes the most nearly middle-valued contour line is selected, tracing it through the field of data points. The cartographer then picks out the data point most appropriate in view of this choice (above), and draws straight lines from that data point to the several (perhaps five to ten) surrounding data points which are nearest. Taking the VAL's of the two data points at the end of each line, he interpolates (and marks the position of) the contour-line values along each line. In Figure 3-19, this process has been performed for two data points, the left-hand one representing an attempt to locate the highest-valued contours (here, the 70-, 60-, and 50-foot contours), and the right-hand one representing an attempt to locate part of the middle-valued contour (here, the 30-foot contour).

Parenthetically, note that only the originally plotted data points can be considered to be "known" points; between these known data points are an infinite number of points, each with determinable LO and LA, but with unknown VAL. If a contour map of the surface under consideration is to be made, this surface must be assumed to be a reasonably smooth, regular



LO (minutes of arc from E041°)	LA (minutes of arc from N17°)	VAL (feet above mean sea level)	LO (minutes of arc from E041°)	LA (minutes of arc from N17°)	VAL (feet above mean sea level)
+19.6	+20.8	+38	+32.0	+17.7	+39
20.0	25.9	25	33.1	22.5	59
21.2	15.0	37	33.9	12.0	33
22.4	20.8	55	34.1	26.0	35
23.1	11.7	35	34.7	22.0	68
23.6	17.9	41	36.0	29.8	15
24.4	24.2	40	37.7	19.8	42
24.5	30.1	14	39.6	22.3	39
25.0	21.1	73	39.8	15.7	30
26.4	19.8	54	40.4	29.9	9
26.6	21.0	61	40.9	14.7	24
27.7	23.1	51	41.4	22.6	32
28.0	26.6	32	42.8	22.8	19
28.7	28.5	21	42.9	12.0	16
29.1	15.5	38	43.5	18.7	17
29.9	21.4	55	44.8	29.2	5
31.4	25.0	39	44.9	23.3	11
31.5	22.9	49	45.7	13.0	8

Figure 3-17-A Land surface, and -*B*, table of data points, used to demonstrate procedures of constructing contour-type maps (Figures 3-18 through 3-23).

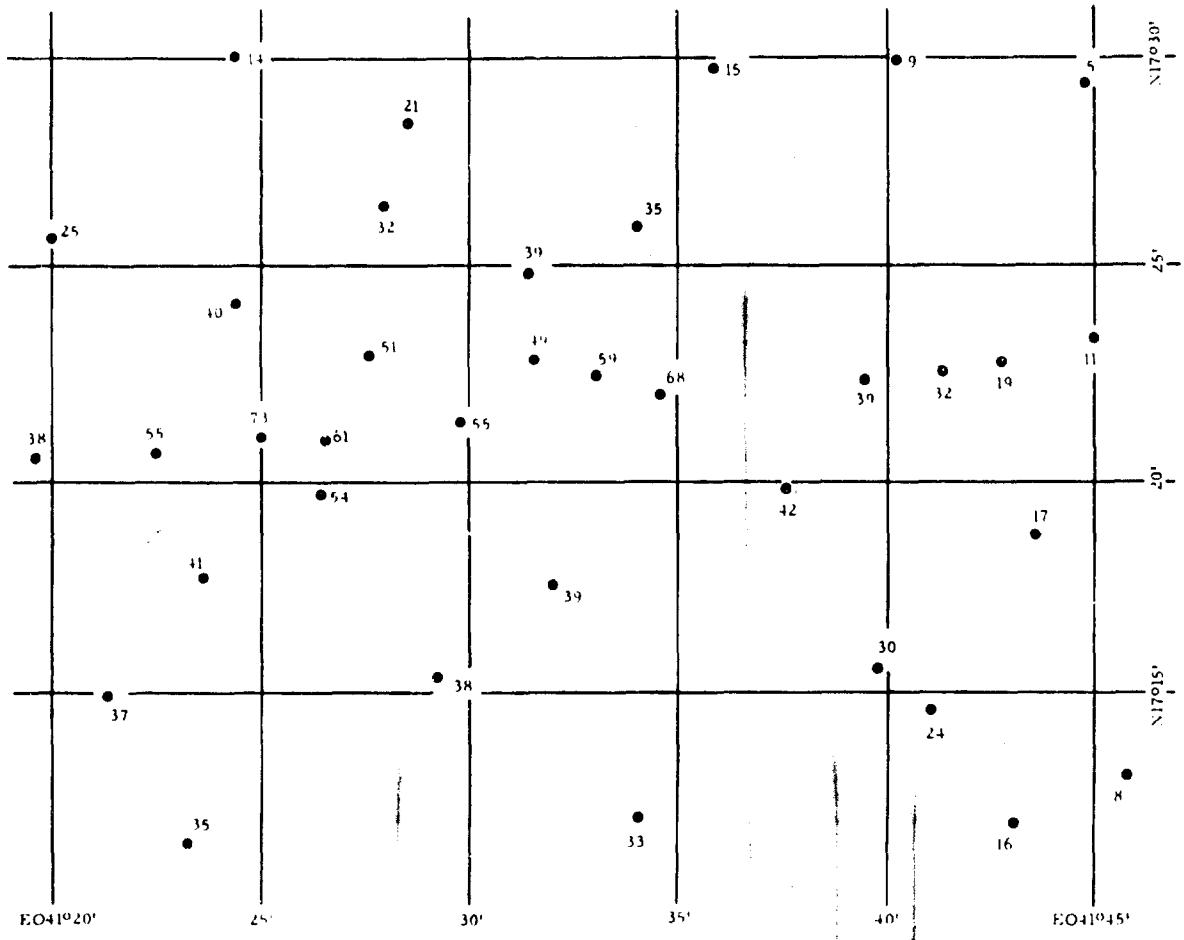


Figure 3-18 Data points (from Figure 3-17-B) plotted on square (LO, LA) grid.

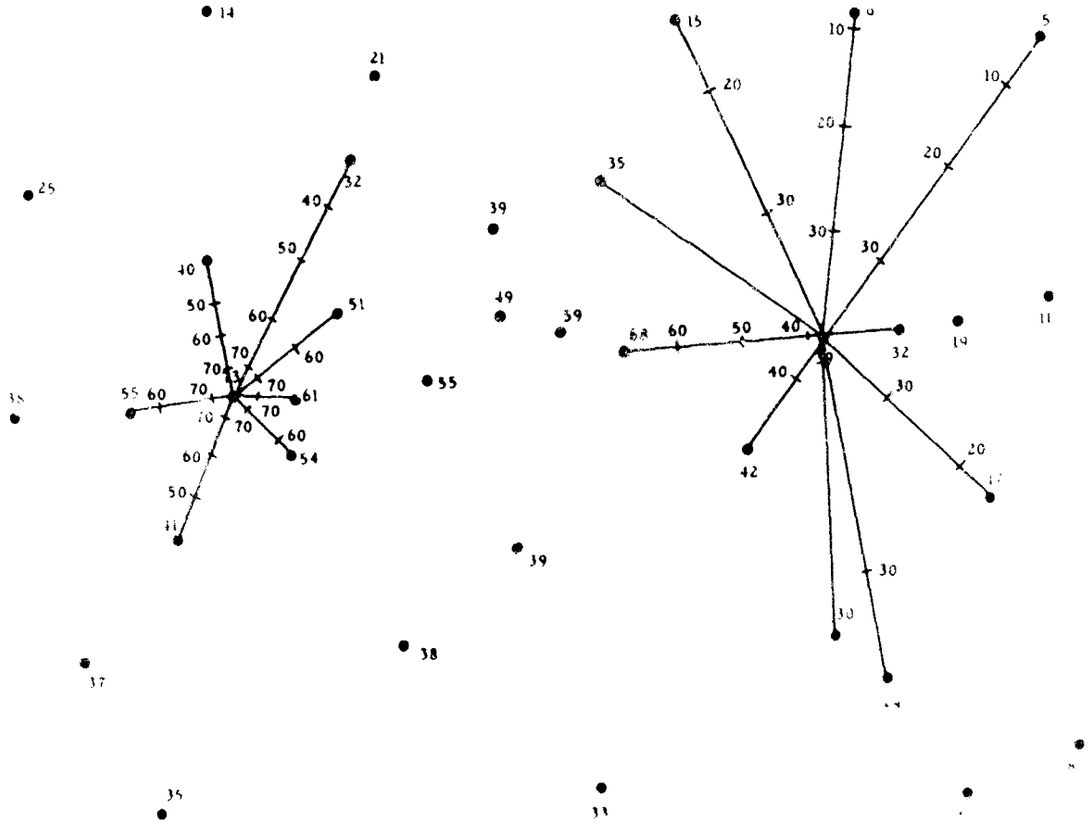


Figure 3-19 Interpolation of contour-line values around two data points.

geometrical surface. Only under these conditions is it meaningful to assign interpolated (inferred) numerical VAL's to the many other loci distributed among the few known data points, and <u>every</u> unknown point must be assigned a mathematically-manipulatable, numerical value (rather than a "?" value).

▶ Then the cartographer continues in this manner, selecting other appropriate (known) data points and interpolating contour-line values around them, building up a large number of rather closely spaced interpolated data points to supplement the known ones. This phase is illustrated by Figure 3-20 in which interpolated data points with VAL's of 70, 60, and 30 feet are indicated. (Interpolated data points with other VAL's have been left off the map for purposes of clarity.)

▶ At this stage the cartographer draws a line connecting successively adjacent, known and/or interpolated data points with identical VAL's, and draws lines in this fashion for each VAL to be contoured. These lines are irregular at this stage and can be considered as <u>preliminary</u> contour lines. The 70-, 60-, and 30-foot preliminary contour lines of our example are shown in Figure 3-21; as before, the other contour lines have been omitted for ease of visualization.

▶ The final step comes when the cartographer (visually) inspects the preliminary contour lines, then smoothes them, exactly as one draws a smooth curve through a set of data points presented on a graph. Assumed (slight) inaccuracies in interpolated data points' values, coupled with the presumed regularity of the surface being contoured, permit this smoothing to be done with no real loss of accuracy. The smoothly curved lines resulting from this last operation can be considered to be <u>final</u> contour lines. The 70-, 60-, and 30-foot final contour lines of our illustrative example are shown in Figure 3-22. Figure 3-23 shows all the final contour lines -- i.e., the completed contour-type map. For comparison, the original land surface from which the data (Fig. 3-17B) were taken is shown in Figure 3-17A.

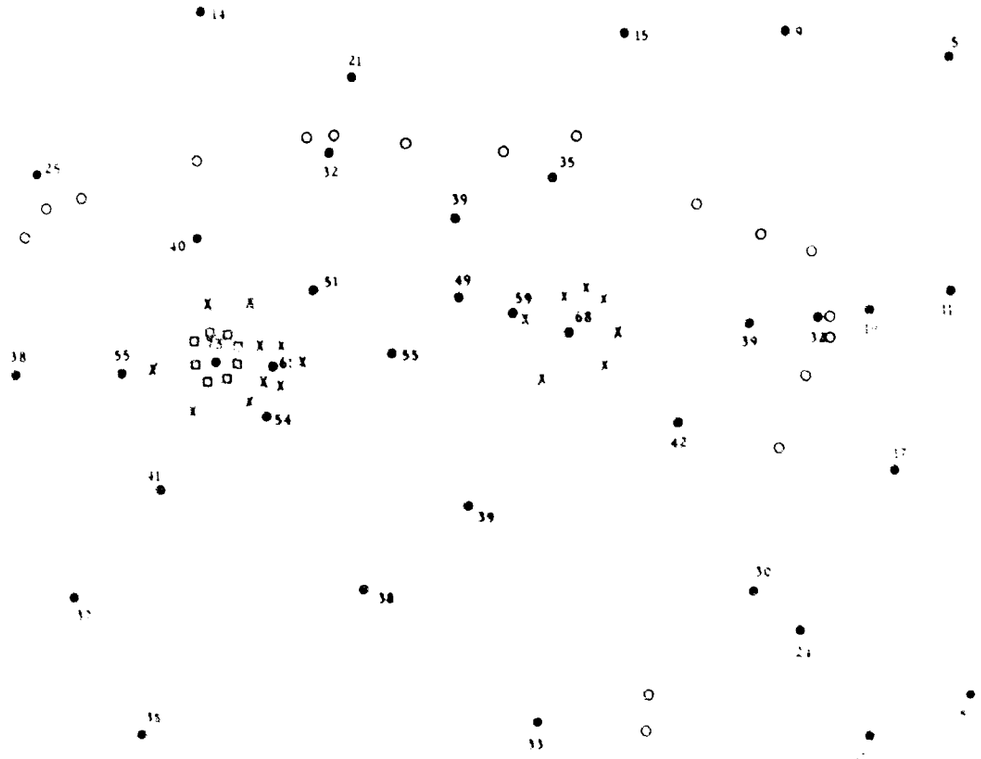


Figure 3-20 Interpolated data points with VAL = 70 feet (□), 60 feet (X), and 30 feet (o).

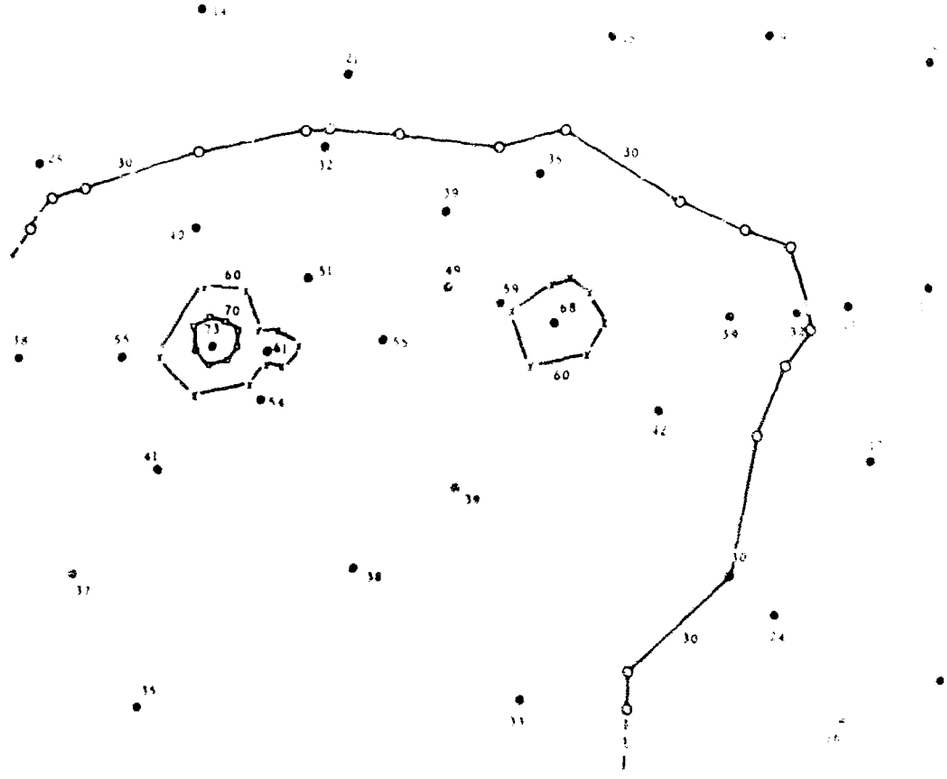


Figure 3-11 Preliminary 70-, 60-, and 30-ppm contour lines.

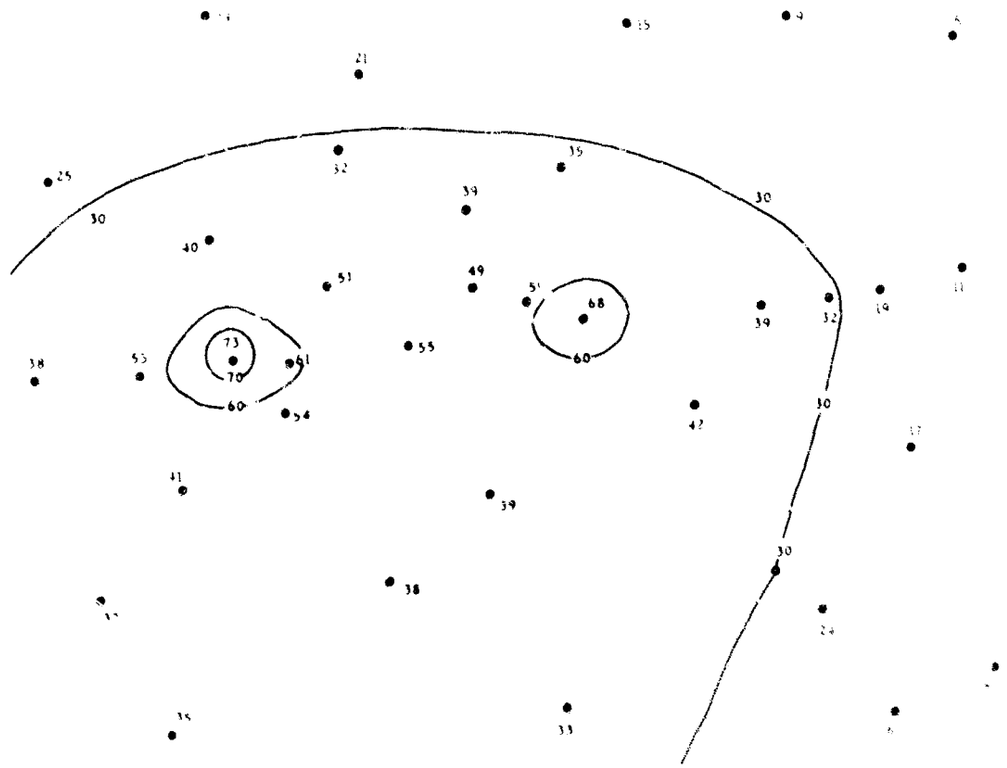


Figure 3-21 Final 70-, 60-, and 30-foot contour lines.

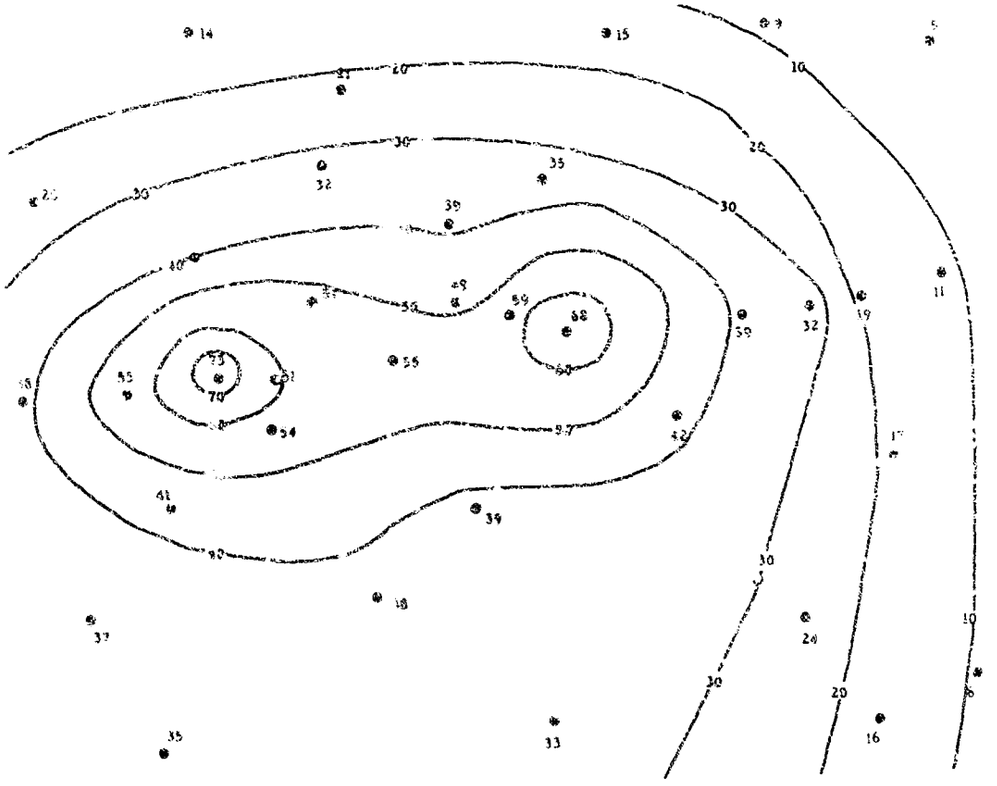


Figure 3-23 Final contour linee drawn from the data points of Figure 3-17-B.

In drawing the preliminary and final contour lines, the cartographer proceeds in accordance with well-defined rules (as outlined by Ohman, 1963). These are illustrated in Figure 3-24.

(1) Most commonly, the contour line will be drawn so as to connect successively the <u>closest</u> geographic point localities that possess the same known or interpolated VAL (Fig. 3-24A).

(2) When points with different VAL's intervene between the two closest points being used to construct a particular contour, the contour line must fit around the first-mentioned points so as to connect points <u>other than</u> the closest points with that particular VAL (Fig. 3-24B). Contour lines should be "smooth" to convey a meaningful concept of the smooth surface which they are attempting to portray.

(3) Contours must be capable of closing either on or off the map (Fig. 3-24C). Contour lines depicted on two separate but adjacent, contiguous maps must connect when the edges of the maps are put together. (Obviously, the form of the land surface does not vary according to how the boundaries of the topographic quadrangle maps showing the land are arranged.)

(4) An elongated ridge crest, representing a reversal of slope or gradient, should be shown by two <u>opposing</u> contour lines of the same VAL, rather than by a single contour line (Fig. 3-24D). This is because, in nature, no ridge crests exist which are of precisely the same elevation (VAL) as a particular contour for a distance great enough to be shown on a map.

(5) When the points being used to draw contour lines are located on the corners of a square, it may be that <u>two</u> possible sets of contours can be drawn for them (Fig. 3-24E). Such a situation can be resolved by interpolating a fifth point at the center of the square with a value (VAL) that is the average of the two possible interpolated values between the pairs of points on opposite corners of the square (Fig. 3-24F).

continued next page

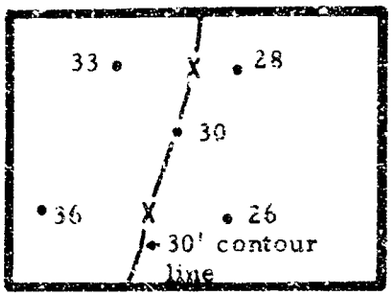


FIGURE A

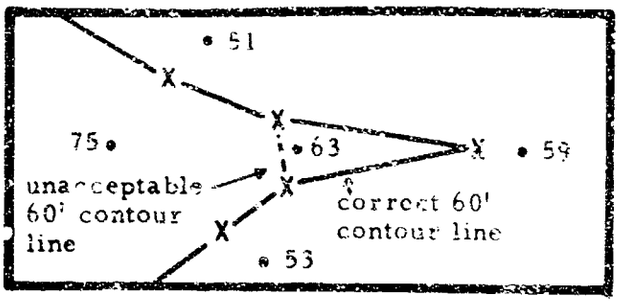


FIGURE B

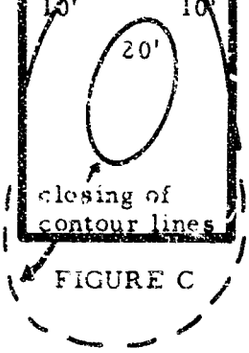


FIGURE C

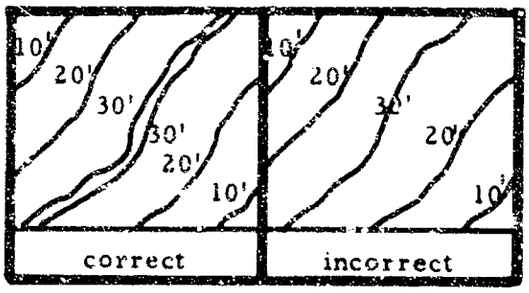


FIGURE D

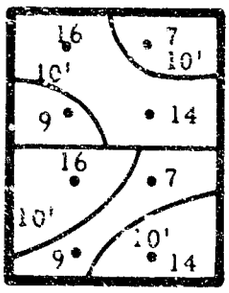


FIGURE E

(• = known points; X = interpolated points)

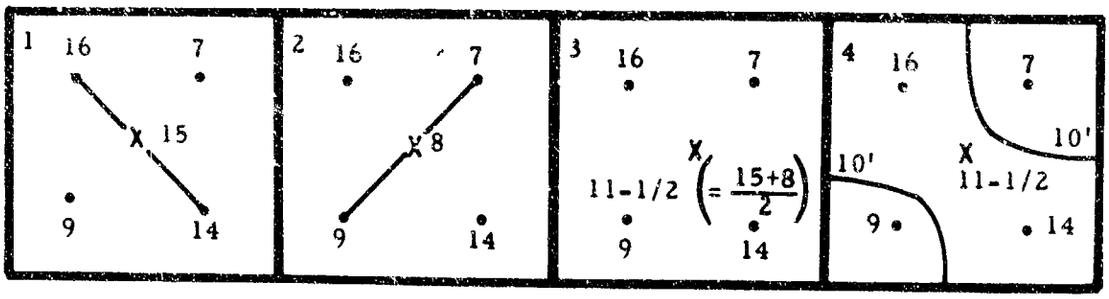


FIGURE F

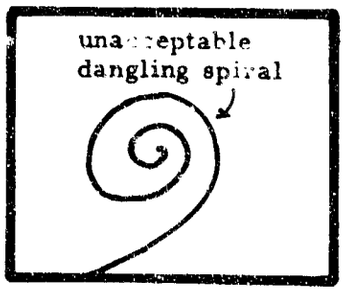


FIGURE G

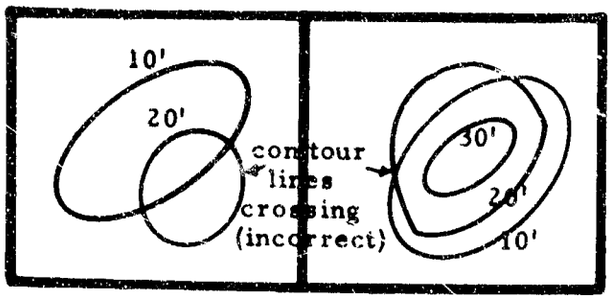


FIGURE H

Figure 3-24 Illustrations of rules to be followed during contouring.

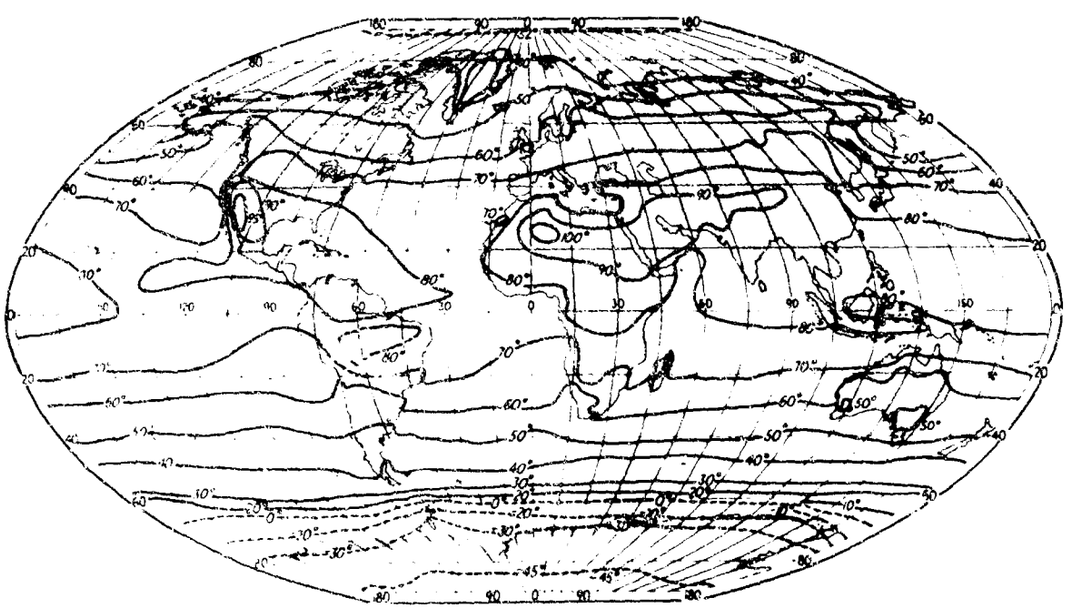
(6) A contour line must never be a dangling spiral (Fig. 3-24G).

(7) Contour lines used to map a particular factor must never cross. Obviously, referring to our topographic map, a particular geographic point cannot be both 10 feet and 20 feet above sea level.

Contour lines greatly aid comprehension in that they tie together points of comparable value and, in addition, show rates of change, thereby stressing relationships which might otherwise be less apparent. Such information is completely lost with shading-type maps since all values in an area are combined to produce an average -- a value which may not even exist, as such, in the area.

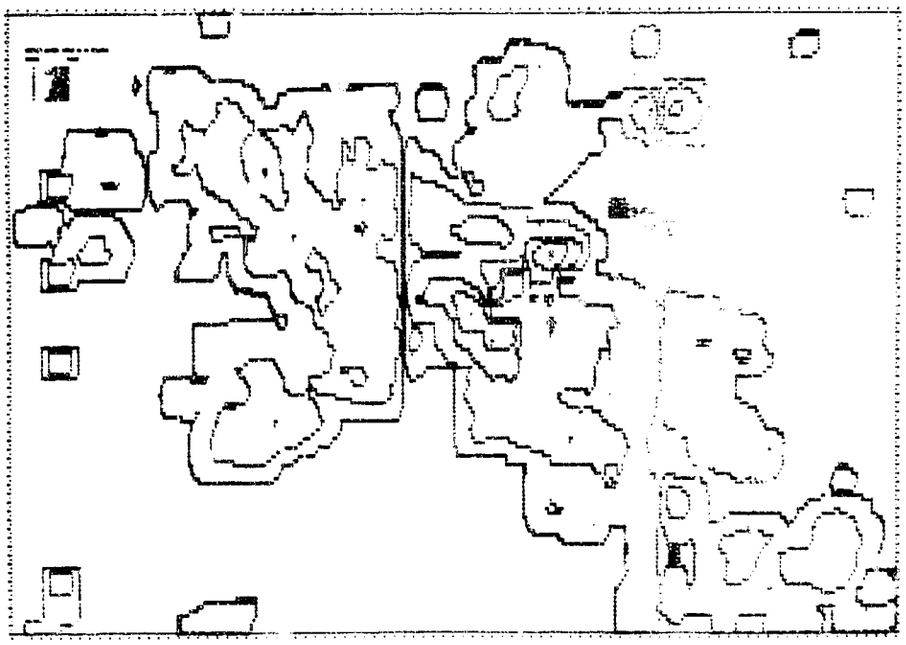
Thus, shading-type maps (and to even greater extent dot-type maps) present quite limited information about disease-environmental relationships because of problems involved in grouping data, problems which often lead to marked distortion of place/quantity relationships. (An exception is the special case where, using high resolution (large scale) maps, every case (or small groups of cases) of a particular disease is precisely located by a dot. This technique is widely used, and very successfully by epidemiologists.) This seems an adequate explanation of why disease-environmental maps have not been more widely used. The method of contour mapping gets around the most serious of the data grouping problems (mentioned above), but makes much greater demands of the data in terms of "completeness".

Various contour-type maps are shown in Figures 3-25 through 3-29, including one that deals with biomedical problems (Fig. 3-27). (There have been relatively few contour maps produced that deal with disease-environmental data. The problems which we have encountered in developing the MGD system, particularly those dealing with data structuring, would explain this.) MOD system produced contour maps, based upon the standard set of (schistosomiasis) data are also shown. Figure 3-28 is one that was drawn manually. Figure 3-29 was produced by a computer. Further examples will be presented later in the discussion of computerized mapping programs.



from Introduction to Meteorology by S. Petterssen, 2nd ed., 1958; reproduced with permission of McGraw-Hill Book Co.

Figure 3-25 Published, manually drawn, contour-type map showing the environmental factor "mean sea-level temperature in July" -- Petterssen, 1958.



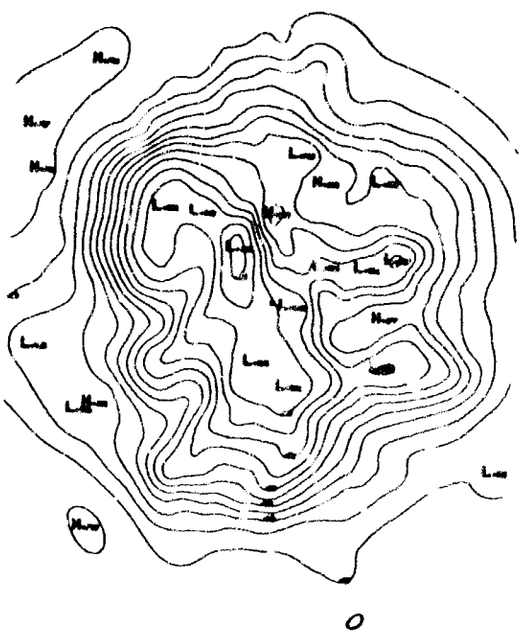


Figure 3-26 Examples of machine-drawn, contour-type maps: A (upper), showing 1960 human population of Ann Arbor, Michigan, output on a line printer (Tobler, 1966, p.7; courtesy of W.R.Tobler,) and B (lower) showing hydrographic data (Tobler, 1964 p. 4; courtesy of W.R.Tobler).

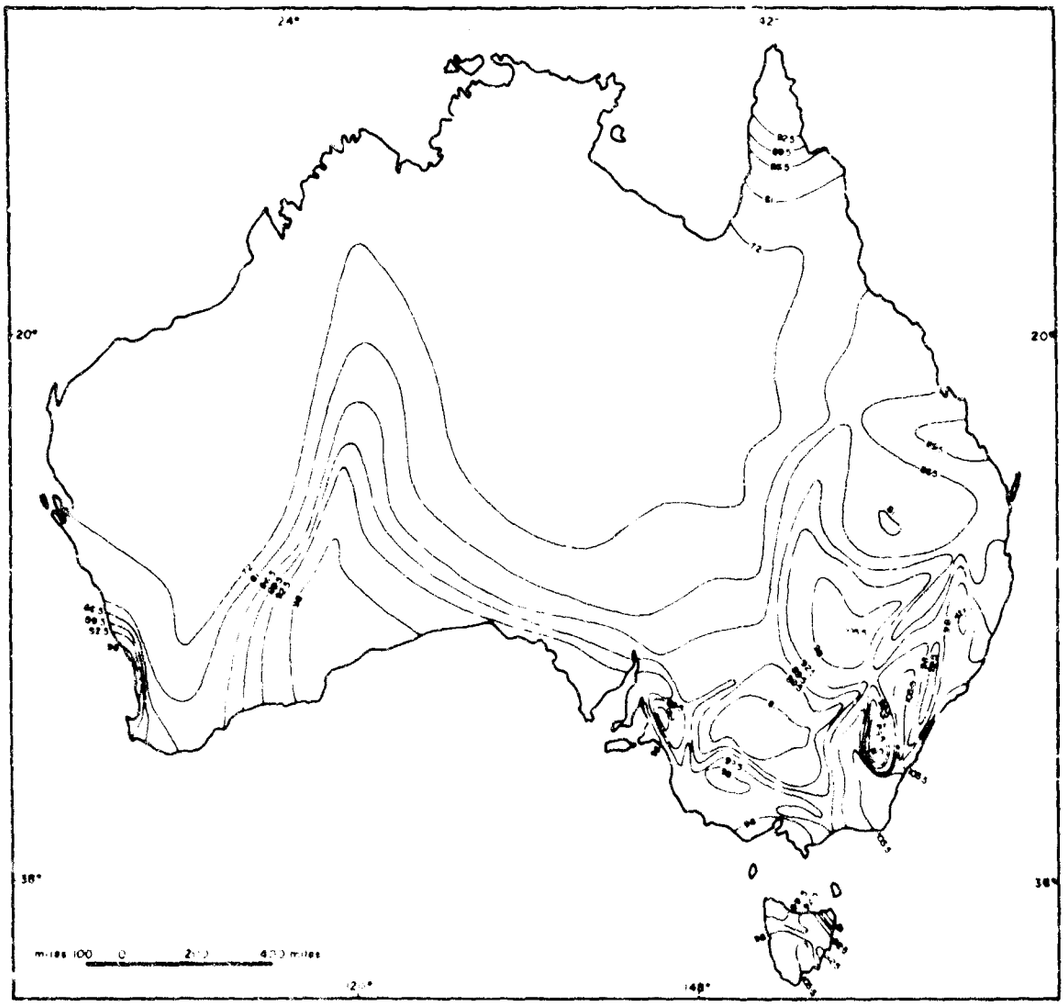


Figure 3-27 Published, manually drawn, contour-type map displaying areal variations of standardized mortality ratios calculated for Australian human males who died from arteriosclerotic and degenerative heart disease during 1959-1963 -- Learmonth & Nichols, 1965.

from Maps of some standardized mortality ratios for Australia, 1959-1963: Occasional Paper No. 3, by Learmonth, A.T.A. and Nichols, G.C., 1965, The Australian Natl. Univ., Canberra; reproduced with permission.



Infection rate (%) of schistosomiasis (*mansoni*) in human beings; data grouped by provinces.

Figure 3-28 Contour-type map of standard schistosomiasis data (Fig. 3-1), drawn manually as part of the MOD study effort.

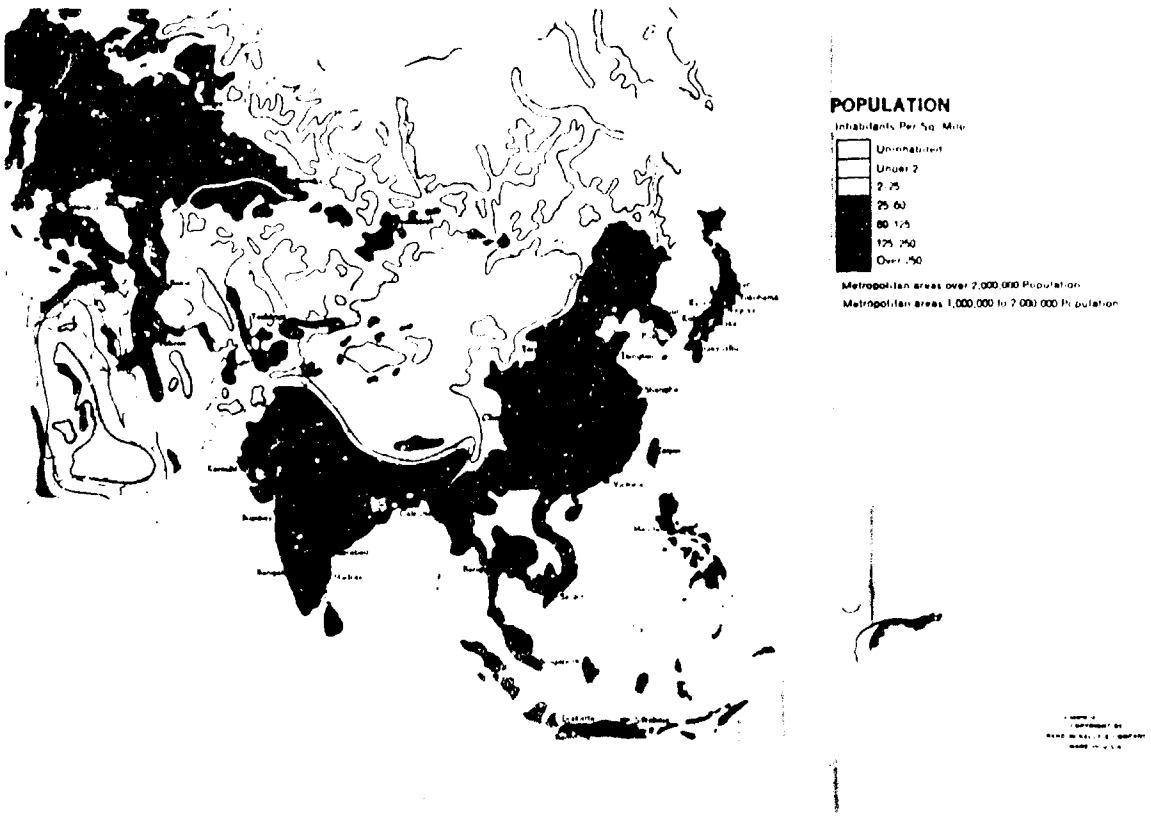


Figure 3-14 Contour map of standard schistosomiasis data (Fig. 3-1), produced by a CDC 3600 computer using Control Data Corporation's gridding/contouring program with a coarse grid, and output on an ink-on-paper CalComp plotter (outline of continent added manually).

3.2.5.4 Combination-type Maps The basic types of maps we have described can be combined to increase understanding of the data. Figures 3-30 through 3-34 show schematically various combinations of dot-, shading-, and contour-type symbols. Figures 3-30 and 3-31 illustrate combination-type maps of particular environmental factors. The standard set of (schistosomiasis) data mapped by the MOD team, using various methods, is shown in Figure 3-33 as a manually-produced map on which a combination of dot-, shading-, and contour-type symbols is used. These data appear again in Figure 3-34 as a computer-produced map utilizing dot- and contour-type techniques.

3.2.6 POSITION, SCALE AND PROJECTION

One of the most important tasks in preparing data for mapping is to determine the best representation of the location (on a LO, LA coordinate basis) for each bit of datum to be mapped. This is because disease factors cannot be measured in the same way as points for a topographic map. In topography, a point location has a single exact value which is directly measurable, and related only to that one location (e.g., 2,398 feet above sea level). Disease infection rate, on the other hand, must be calculated for an area by summing the total number of animals (including human beings) in the area and dividing this by the number of infected animals (of the same species or group) in the area. The value thus derived is absolute only with reference to the area boundaries, and would change if area boundaries changed (even though the same point location might be used for the data location in each case). Difficulties often occur when data from different size areas are combined on one map. The larger the area considered, the less variation would show on the map. To carry this to an extreme, the area covered by the whole map could be represented by one point. This difficulty often occurs in relation to politically defined regions. If averaged data for one country were used to represent that area, all other data points on the same map should also be grouped by country; data from unique villages or rural areas would be useful in making other maps, but



from Goode's World Atlas, 12th ed., 1964
Copyright by Rand McNally & Co., R.L. 68 S 86; used with permission.

Figure 3-30 Published map combining contour-type (the lines separating differently colored regions) and shading-type (the colors between the contours reproduced here as different shades of gray) symbols showing population data.

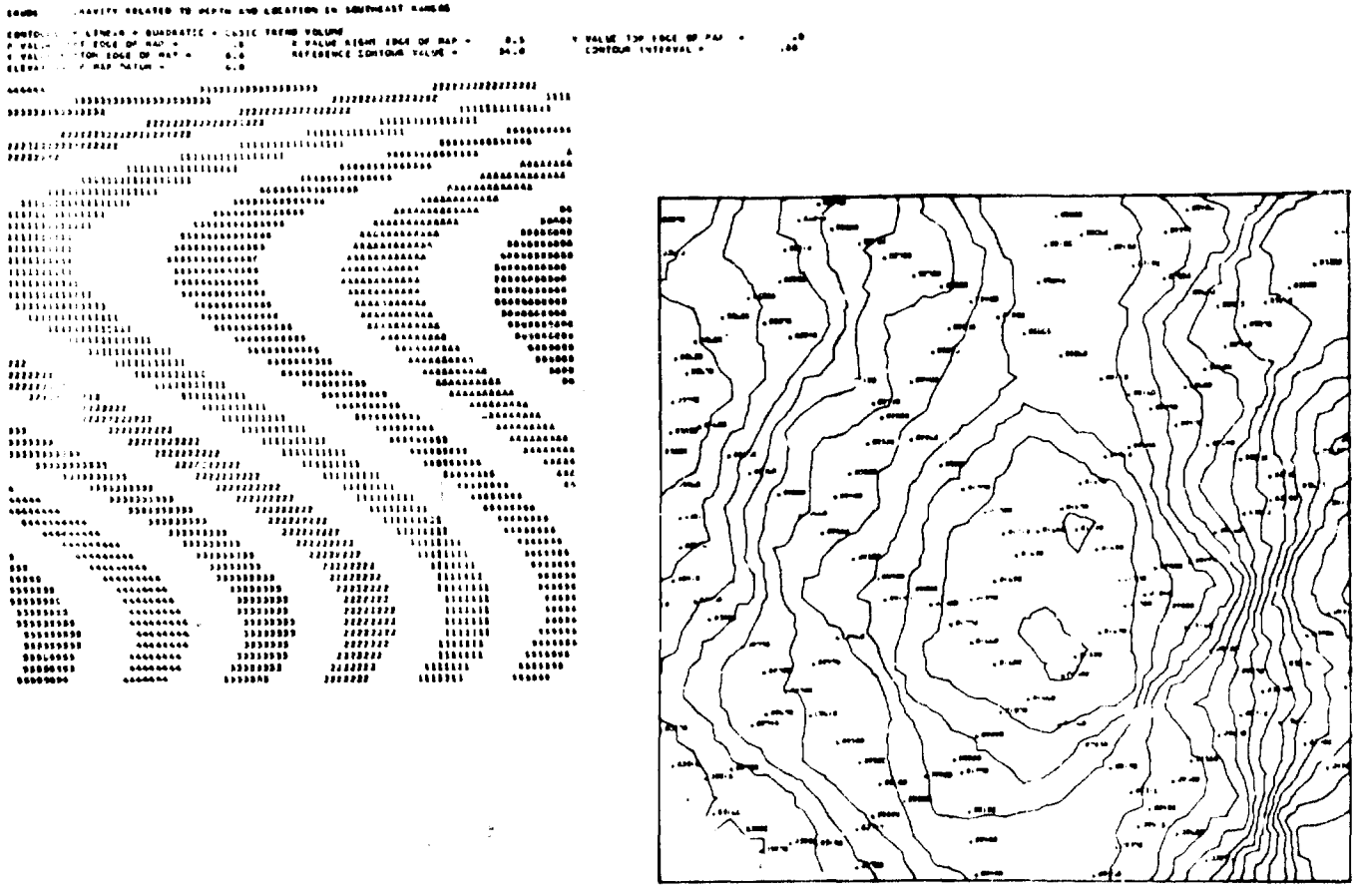


Figure 3-31 Machine-computer-drawn combination-type maps: A (upper), utilizing combination of shading-type (the alternating bands of characters) and contour-type (the boundaries between adjacent white and black bands) symbols to portray crude oil gravity data, output on line-printer (Harbaugh, 1964, p. 56; courtesy of State Geological Survey of Kansas); and B (lower), utilizing contour-type and dot-type symbols to portray hydrographic data, output on a CalComp plotter (courtesy of California Computer Products, Inc.).

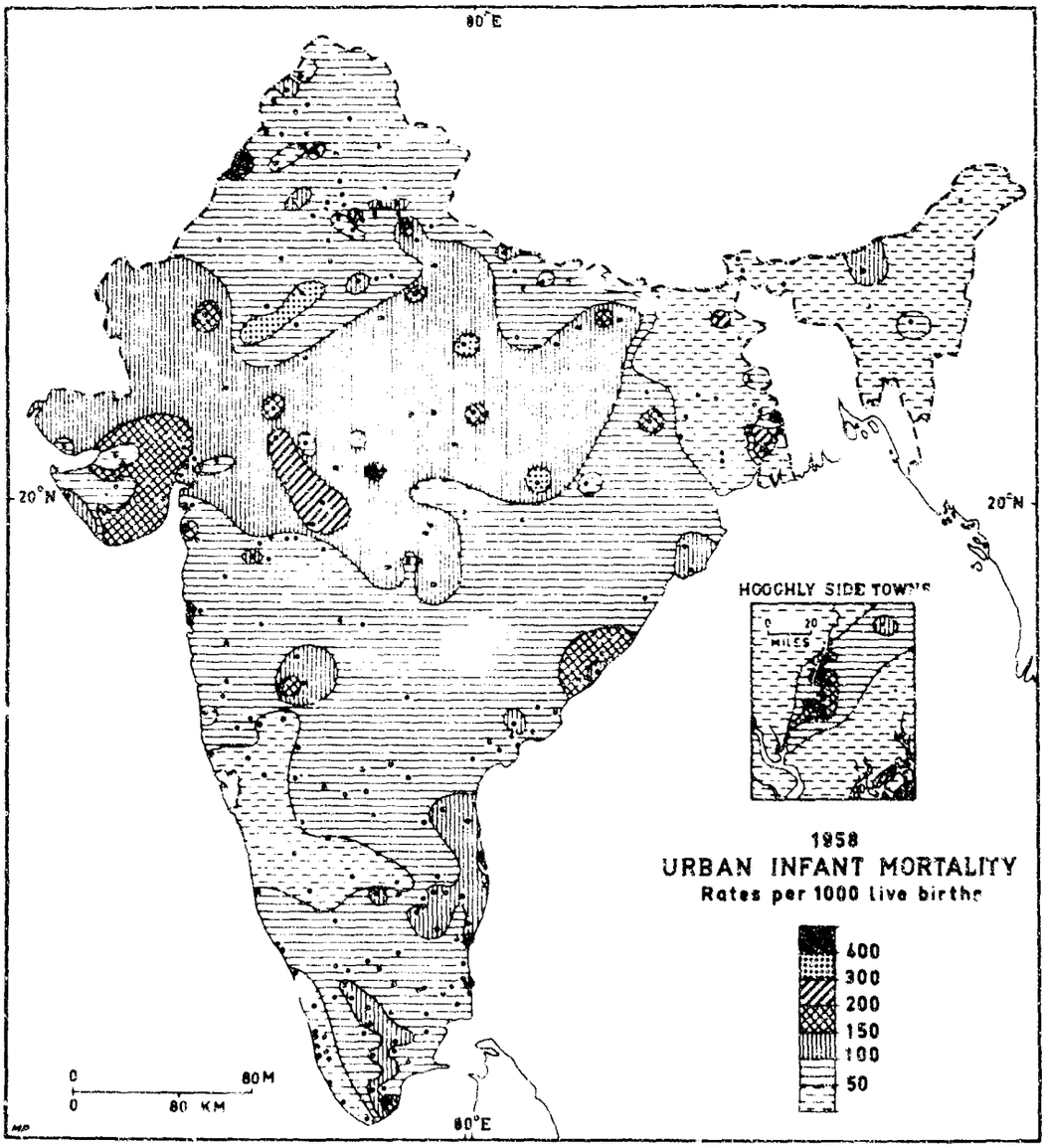


Figure 3-38 Published, manually drawn map using combination of dot-, shading-, and contour-type symbols to portray urban infant mortality (number of deaths per 1000 live births) in India in 1958 -- Learmonth, 1965.

from Health in the Indian Subcontinent, 1955-1964: Occasional Paper No. 2, by Learmonth, A.T.A., 1965, The Australian Natl. Univ., Canberra; reproduced with permission.

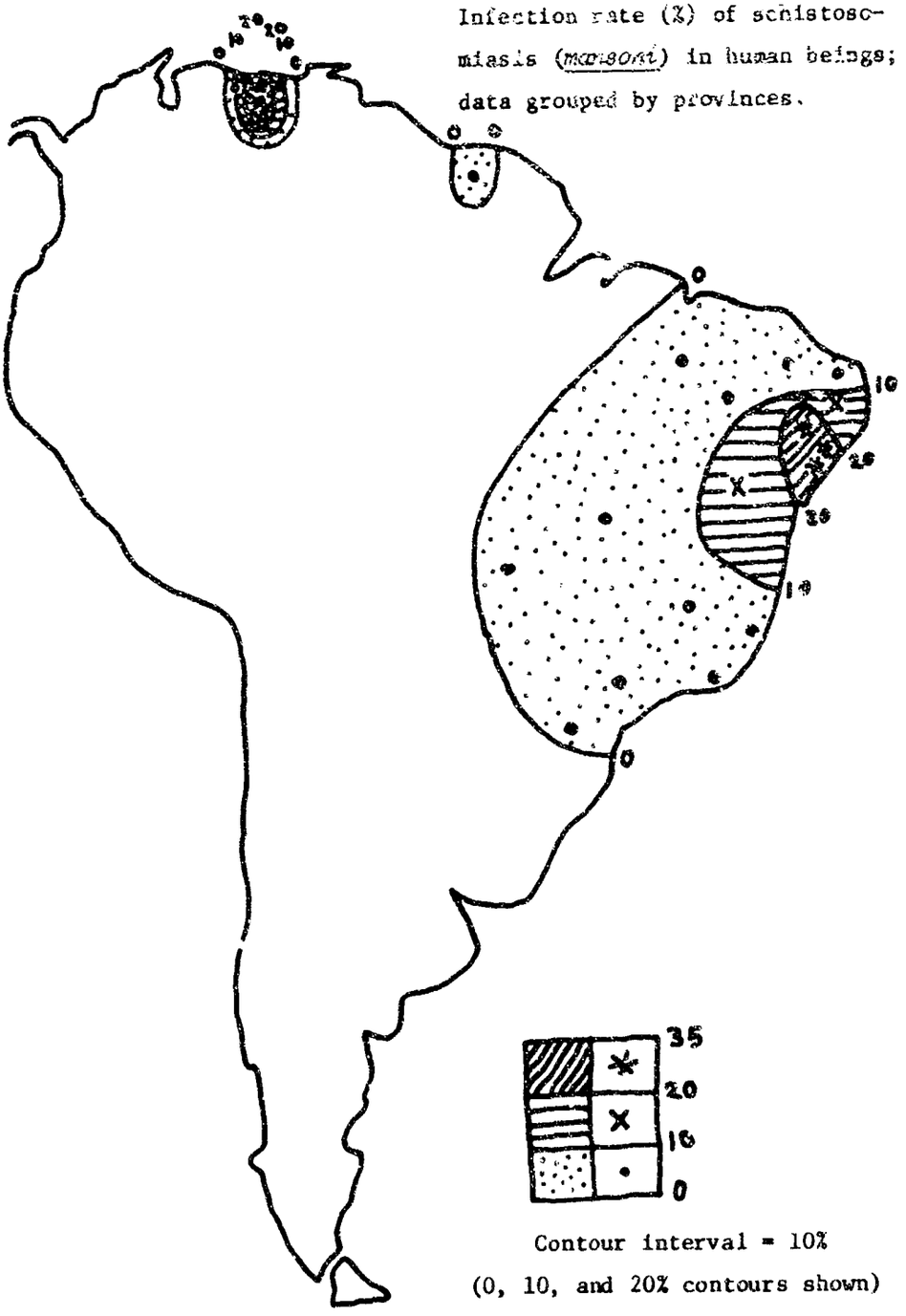


Figure 3-3. The standard MOD set of schistosomiasis data (Fig. 3-1) presented as a manually drawn map utilizing dot-, shading-, and contour-mapping techniques.

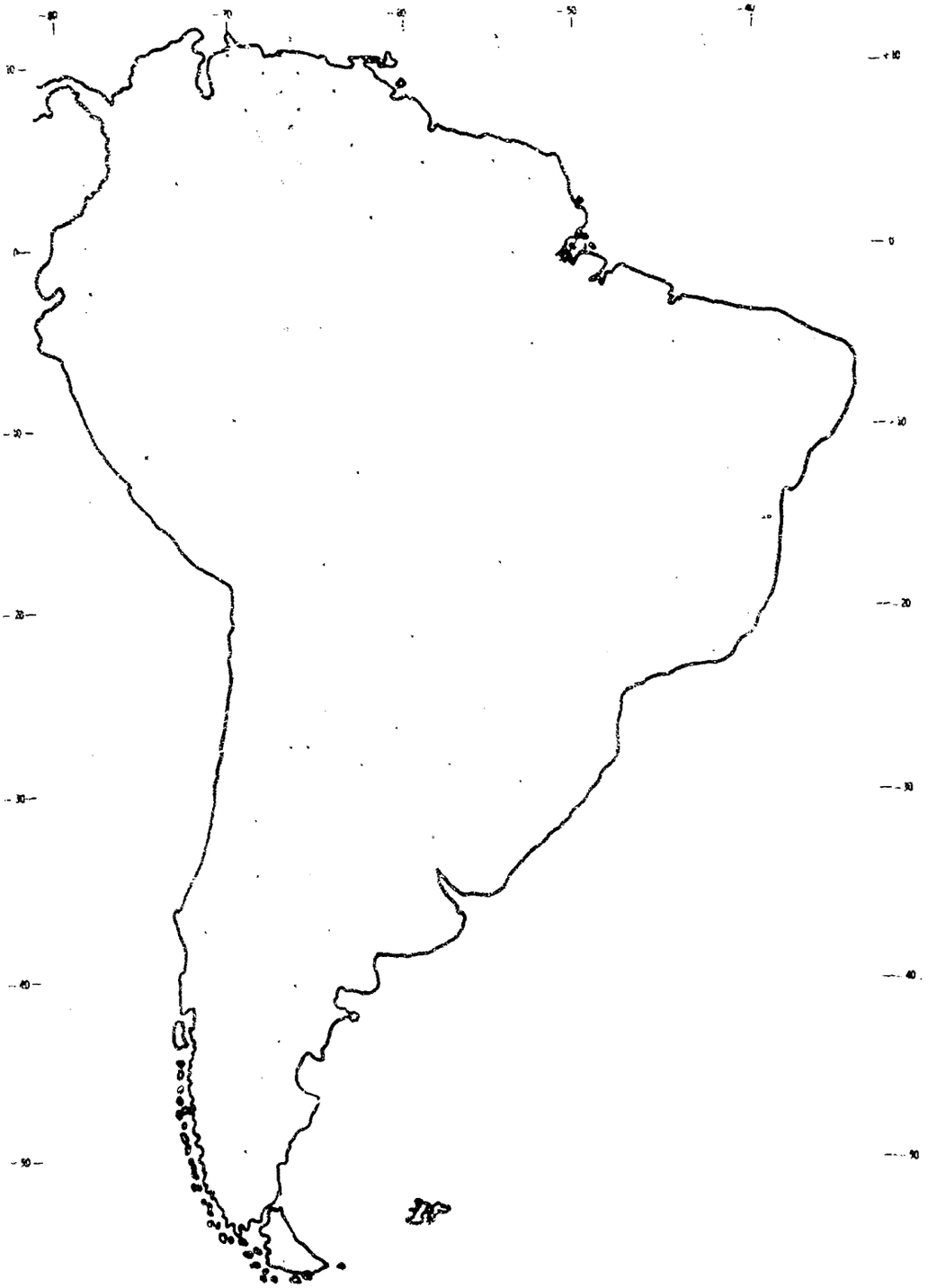
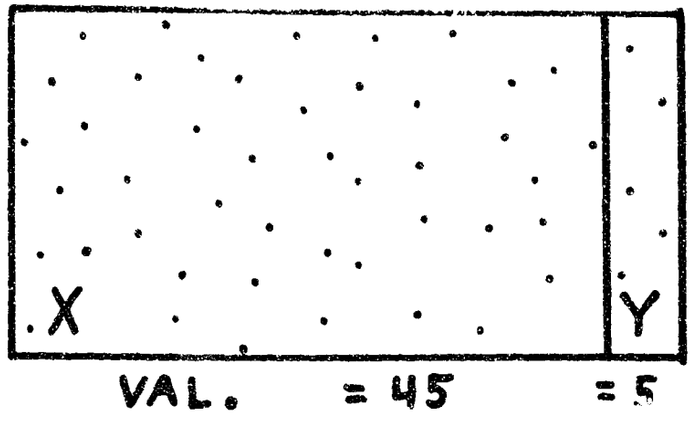


Figure 3-34 The standard set of South American schistosomiasis data (Fig. 3-1), presented as a combined dot-type (the X's) and contour-type map, drawn by a CDC 3600 computer utilizing an offline ink-on-paper CalComp plotter and the Control Data Corporation's gridding/contouring program with a fine grid (outline of continent added manually).

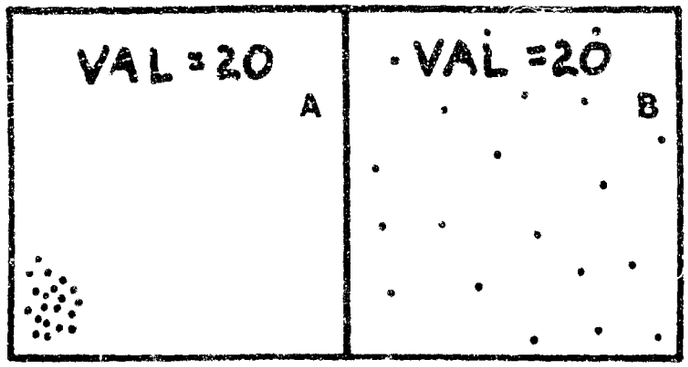
should not be used in combination with country grouping. When data are grouped in an artificial manner such as this, the user must be aware of their limitations. Disease is no respector of political boundaries, and disease data that are grouped (and mapped) by country or province (often a technical necessity) may be quite misleading. For example, knowing that a county has a 25% overall infection rate of a specific disease would not indicate that a particular village within that county had an 80% infection rate whereas the surrounding rural areas had an infection rate of less than 1%. The illustrations on the following two (insert) pages, 3 - 61A and 3 - 61B, emphasize some of these limitations.

Despite the limitations we have mentioned, data must often be grouped and related to areas larger than desired, and we must do the best we can with them. One of the many possible methods of handling such a problem would be to locate the grouped data at a "center of gravity" based upon population distribution. Ideally the MOD system would use regularly shaped areas, preferably square, forming a grid, as a basis for map production -- and the grid squares would be adjusted (in size) as necessary to reflect a reasonably uniform disease-environmental situation. All data within a given square could be combined at the center by calculating an inverse distance function for each datum value. Squares with no data could, if desired, be exempted from further consideration in mapping. Unfortunately, this concept is not universally accepted because, in non-computerized mapping, gridding data points is considered to be unnecessary (as we discussed in 3.2.5.1 in relation to dot-type maps).

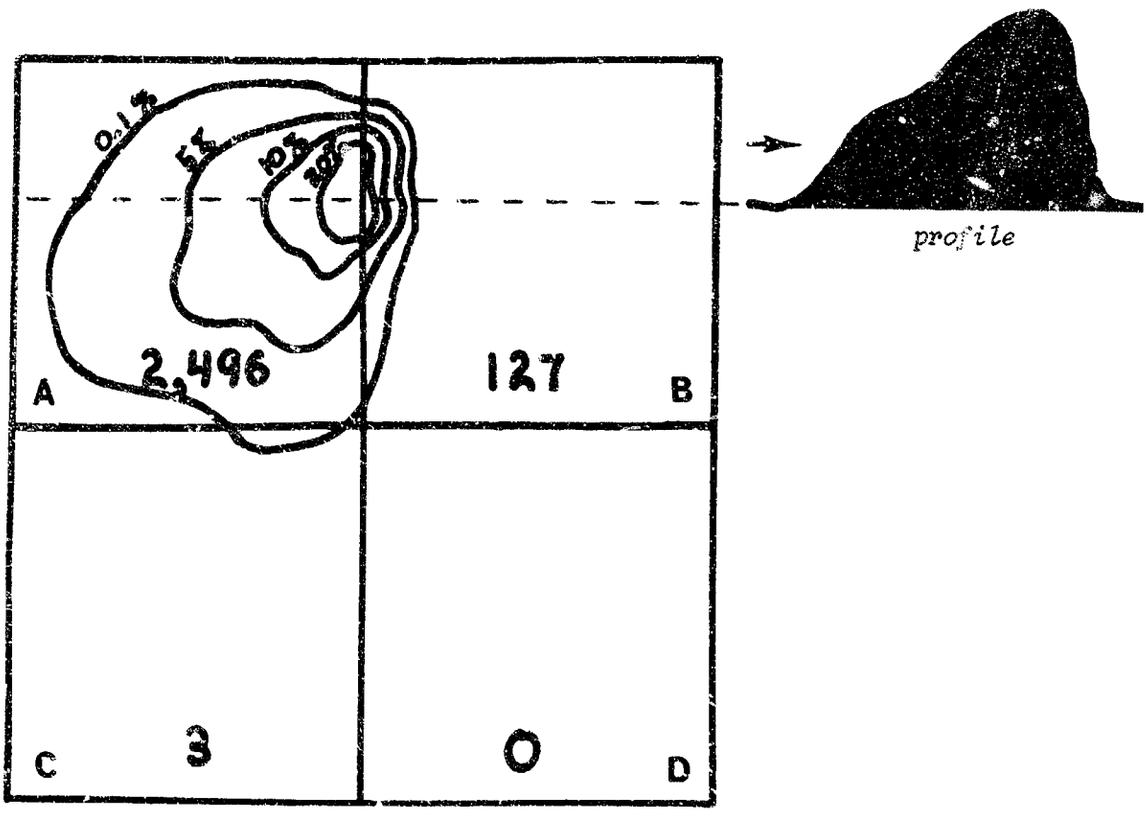
Map scale and map projection warrant consideration before we describe the actual construction of disease maps. Map scale is expressed in three ways: verbally (e.g., one inch equals 40 miles), graphically, by a scale which is drawn on the map -- a line divided into units which represent actual distances on the map, and fractionally, where the scale is indicated by a ratio, e.g., 1/25,000. Map scale is determined by the area to be mapped and the desired dimensions of the map. These determinations are, in turn,



(A) An illustration of the misleading information that can come when disease data is expressed in relation to political (unit) areas, without consideration of actual (mathematical) areas.

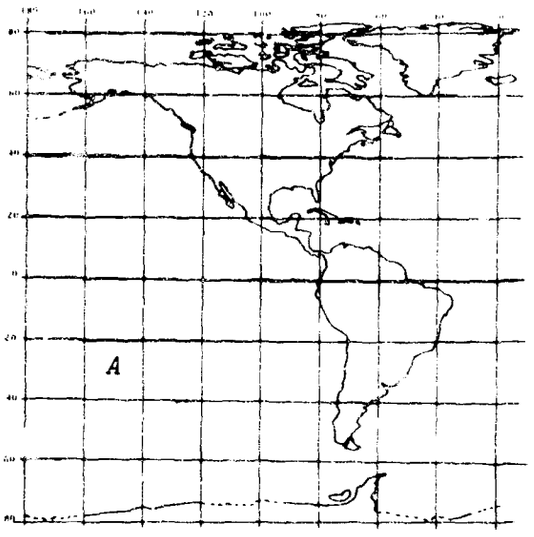


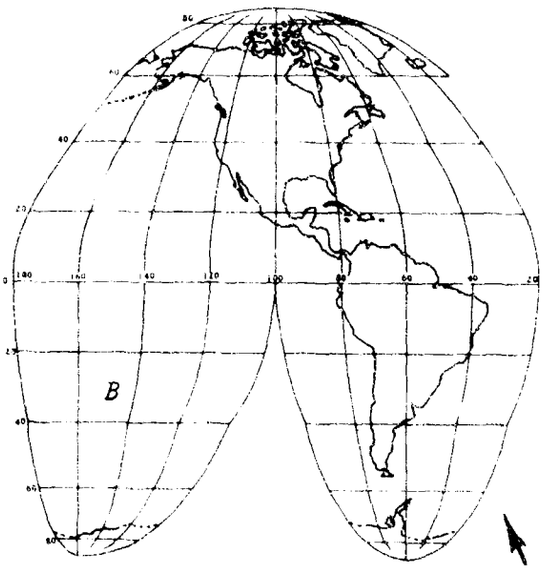
(B) An illustration of how disease prevalence or incidence figures can be very misleading when related to political (unit) areas, even though they may be of equal size.



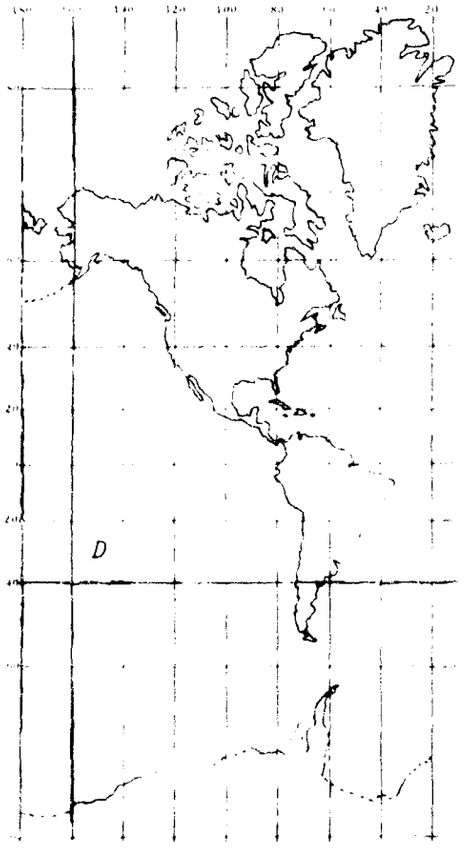
(C) An illustration of how one might form a false impression from disease data reported by political areas. In this example the disease is altitude dependent, and its distribution related to a mountain, not to the artificial boundaries of Provinces A, or B, or C, or D. Although we have not shown it here, this altitude relationship would be immediately apparent upon overlaying the (transparent) disease map on a base topographic map.

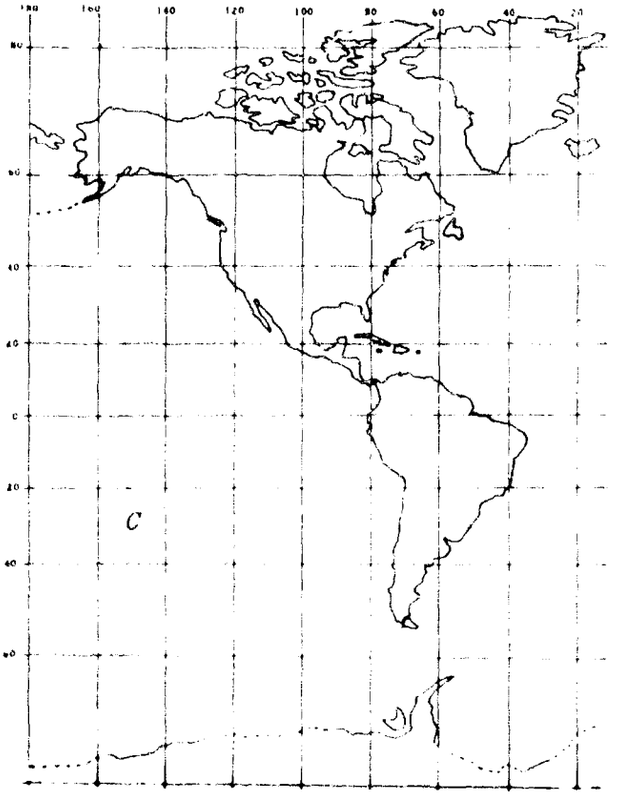
Obviously, disease prevalence or incidence data, when presented in dot-type or contour map form -- as shown in figures A, B, and C -- would not be subject to the type of misinterpretation that could arise if the data were reported simply as figures, related to political areas.





Copyright by the University of Chicago; reproduced with permission.





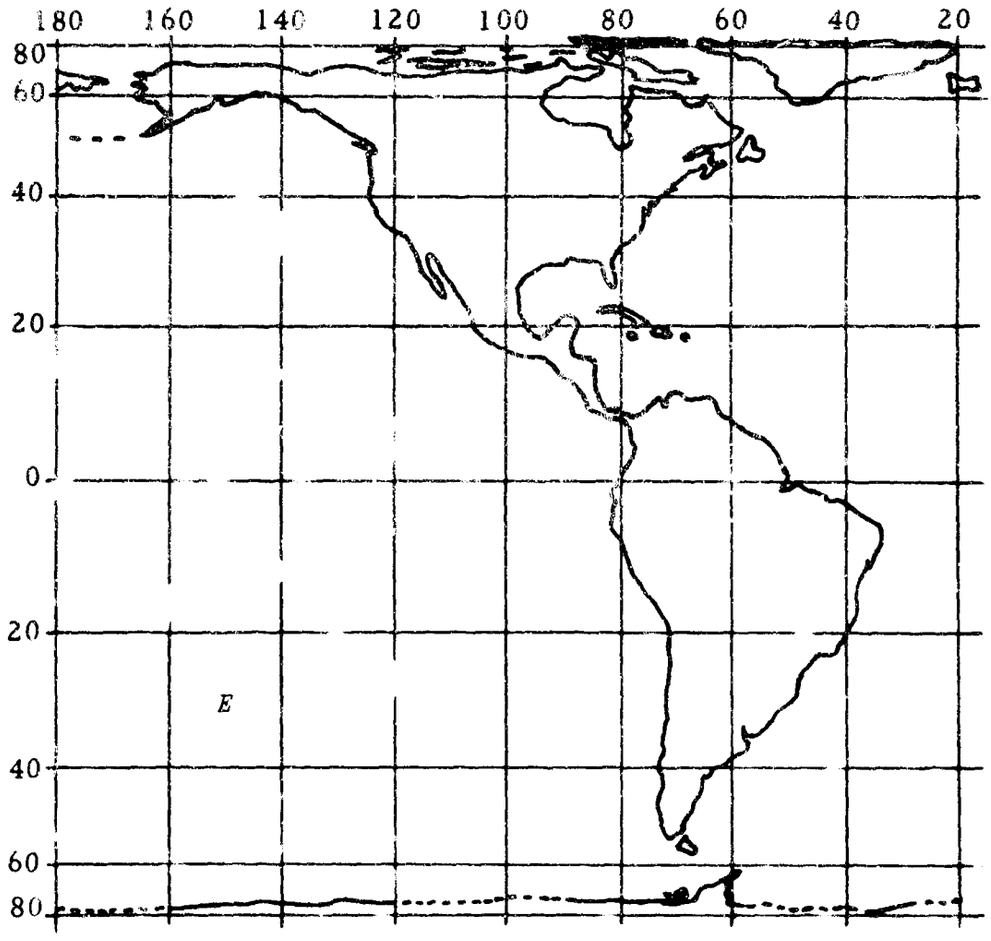


Figure 3-35 The Western Hemisphere mapped on map projections found to be most useful to the MOD system: A, equirectangular projection, B, Goode's homolosine projection, C, Miller cylindrical projection, D, Mercator projection, E, cylindrical equal-area projection. Note that in E, distortion is minimal in the tropical zone.

from Elements of Cartography, 2nd ed., by Robinson, A. H., 1960, published by John Wiley and Sons, Inc., New York and reproduced with permission.

influenced by whether or not the dimensions of the desired map exceed the limitations of paper size and how many points will probably be used. For example, if the data were recorded only to the nearest ten miles, a large scale (small area) map could contain very large errors. As the size of the area covered on a map becomes smaller, the scale increases and "resolution" increases; conversely, the larger the area covered, the smaller the scale, and "resolution" decreases. These are, obviously, very practical considerations. For example, it would contribute very little to use (small area) local data on a map of the world unless it were first combined with data in adjacent areas.

Map projection also poses several problems. No one map projection is best for all purposes. Each type of projection shows certain characteristics of the earth better than do other projections, and each projection distorts some characteristics of the earth's surface. Distortions of size and of shape are two major kinds of effect. Since MOD users will probably be comparing distribution patterns, the best projection (ordinarily) is one that minimizes distortion of areas and shapes (with less concern for linear distances, angles, and directions).

Using these criteria, the homolosine projection (Fig. 3-35B) is probably the most suitable one for maps of the world as a whole. However, two other projections should be considered; the Mercator projection (Fig. 3-35D) and the similar-appearing Miller cylindrical projection (Fig. 3-35C). Even though both distort area and shape rather markedly, they are widely used and, thus, familiar projections, furthermore, much data are already available, mapped according to these projections. The MOD computer-produced maps in this report were all produced at varying scales, but with no manipulations of LO, LA (i.e., LO = X and LA = Y), and can be considered as examples of equirectangular projections, see Fig. 3-35A. The choice of a suitable projection for areas smaller than a continent does not present as great a problem as it does for the world as a whole since all projections tend toward the equirectangular, as a limit, as the region mapped becomes smaller.

The cylindrical equal-area projection seems a good general compromise for a standard MOD projection. It is the only equal area projection with a rectangular grid (see Fig. 3-35E). While it looks unusual when standard parallels just under 30° are chosen, it provides the least mean deformation of any equal-area world projection (Robinson, 1960, p. 75).

3.2.7 MAP CONSTRUCTION

We have considered in some detail the characteristics of maps and the methods of mapping. As we have illustrated, disease data, once collected and structured, can be mapped by conventional (manual) methods, using these same standard cartographic techniques. Let us now consider the methods by which computers produce maps.

There are no new cartographic principles involved in the production of maps by computer, and we are still limited to three basically different ways of representing the data; using dot-type (i.e., data-point) symbols, or shading-type symbols, or contour-type symbols. These three types of symbols can be inserted on a map by computer, using several different output devices (and these devices are described in 6.1.1):

- By high-speed (line-) printers,
- By ink-on-paper automatic plotters,
- By cathode-ray-tube (CRT) devices, displayed directly for viewing and/or recorded on film.

The ways that data symbols are mechanically inserted is mainly a technical problem; not so the construction of data points and the determination of how many and where to place them. Because disease-environmental data are (qualitatively) quite different from the type of data which have been computer mapped, and because of (quantitative) limitations in the number of disease-environmental data points available in many instances, i.e., their sparsity, it was necessary for us to carry out many preliminary exercises - some manually, some with the aid of computers -- to get a clear understanding of the problems.

Dot-type (data-point) mapping techniques were found to be directly applicable to computerization; furthermore, we found that these techniques can often be more easily and accurately performed by a computer than by a person.

Shading-type mapping techniques cannot be carried out by a computer unless the boundaries of the shaded area are completely defined, or unless the entire earth's surface is divided into grid boxes and each box identified. Figure 3-12B shows a successful attempt at shading in connection with the identification of areas by their skeletons. This method (Pfaltz and Rosenfeld, 1967) employs a grid, but it is not required that all grid points be identified, thus computer storage space requirements are minimized.

Figure 3-36 illustrates a technique which, though performed manually by the study team, could be computerized. An alternate method to that used in producing the map of Figure 3-12B is also demonstrated for comparison -- one that utilizes larger grids to reduce further the computer storage space requirement. Figure 3-36A shows the geographic plitical unit boundaries and the data values grouped by province (data taken from Fig. 3-1). Figure 3-36B simulates a computer/plotter produced shaded map of these data. This technique uses shading patterns that can be represented entirely within a single grid box, of such a nature that when grid boxes are combined, the resulting pattern is smooth and uniform. These shading patterns can be lines, or dots, etc. (see Figs. 3-37A through E) combined to form the overall patterns, as shown in Figures 3-37F through I. An alternative technique could be employed directly by a computer and line-printer to produce a sort of shaded map; the character positions of the printer would represent a rectangular grid (with boxes 1/8 inch by 1/10 inch), and the actual print characters would represent the shading symbols. Figure 3-36C demonstrates a simulation of this technique, using the same data. We wrote a simple computer program to demonstrate more effectively this method (illustrated in Fig. 3-36C), using the standard (schistosomiasis)

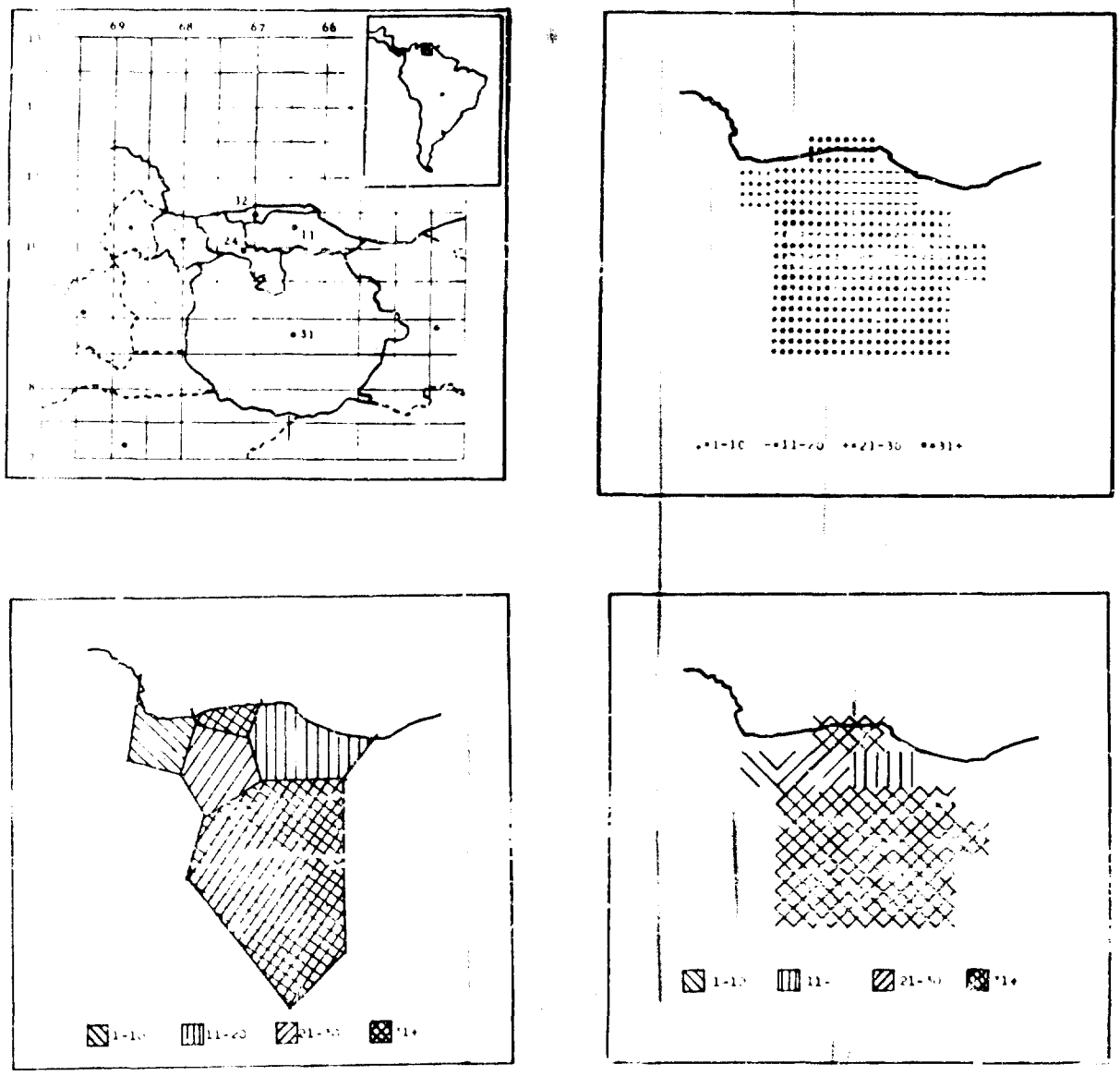
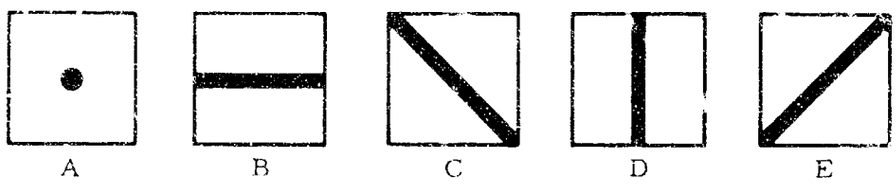
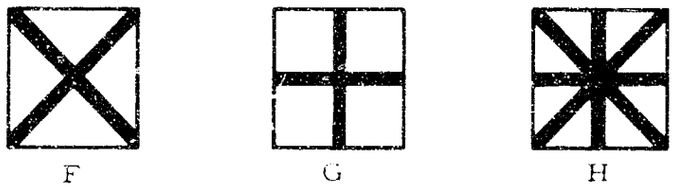


Figure 3-6 A, Standard (Venezuelan) schistosomiasis data (Fig. 3-1), provincial boundaries overlaid with a 1/2° X 1/2° grid; B, shaded to simulate computer/plotter output, based upon 1/2° grid; C, shaded to simulate computer/line-printer output, based upon 1/2° grid; D, shaded to simulate SYMAP program (Fisher et al, 1967). B, C, and D present the same data shown in A.





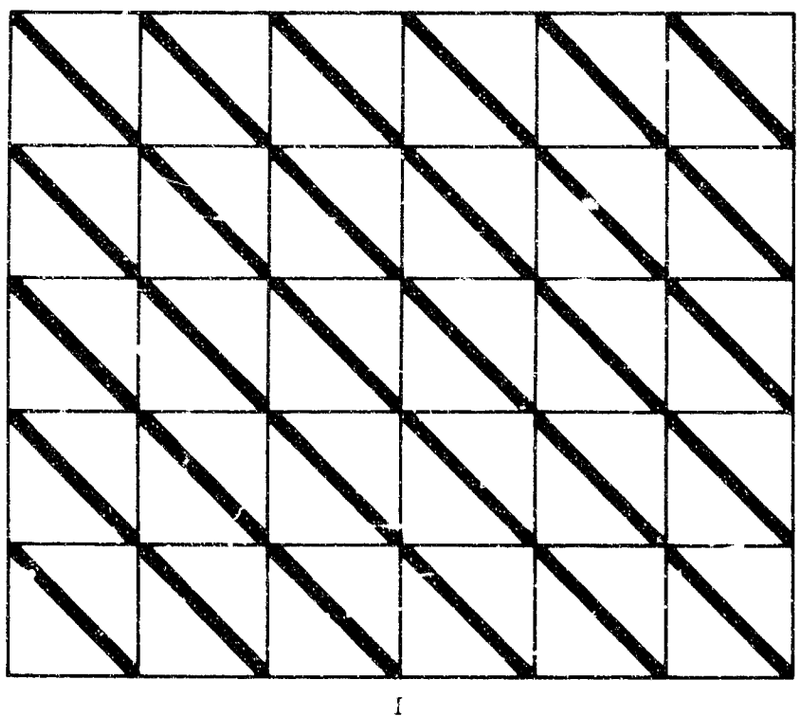


Figure 3-37 Shading patterns contained within single grid box; *A-E*, single grid boxes containing simple characters to be printed or plotted within the grid boxes; *F-H*, single grid boxes containing characters formed by overprinting two or more of the simple characters in A-E; *I*, uniformly shaded area in which the simple characters in each grid box would combine to give the impression of diagonal rulings.

data set. The results are presented in Figures 3-15 and 3-38. Since the data were entered on an equirectangular grid, the resulting map is somewhat unconventional -- an elongated rectangular projection -- but a useful result, nevertheless. Additional (schistosomiasis) data were obtained from the medical literature to extend our preliminary exercises, including consideration of a new geographic area. Approximately 75 data points were plotted on an existing base map of Africa. These data represented reports from a combination of cities, provinces, and regions derived from information presented by Malek in May, 1961. Due to the limitations of such data (related to grouping), the maps that we have constructed do not present disease situations exactly. As we have emphasized before, our primary objective with these various limited disease-environmental data has been to explore techniques of manipulating data to produce mappable information. The geographic location chosen for each of the points was interpolated manually as each value was placed on the map. Then these data were manually contoured by members of the MOD study team. The results are shown as Figure 3-39. Figure 3-39A was contoured using conventional cartographic techniques; Figure 3-39B was contoured essentially the same way, except for using a computer-like method of interpolation. Although there is considerable difference in overall appearance (note the smoother flowing lines in Fig. 3-39A), the maps show definite similarities in the areas where data points are densest. Where data points are close together and relatively uniformly distributed over a broad region, contour mapping is comparatively straightforward, and the resulting contours are [illegible] in serious error. However, where data points are few and far between, problems arise (as illustrated in 3-39B) because it must be assumed that the data pattern being mapped represents a smooth, homogeneous statistical surface. Obviously, this is not necessarily a valid assumption.* Some

*The situation is improved if one has a general idea of the nature of the region being mapped. For example, given spaced elevation point-readings in a particular area, one would use quite a different approach to interpolating if he knew that he were dealing with midtown Manhattan as compared with Kentucky hill country.

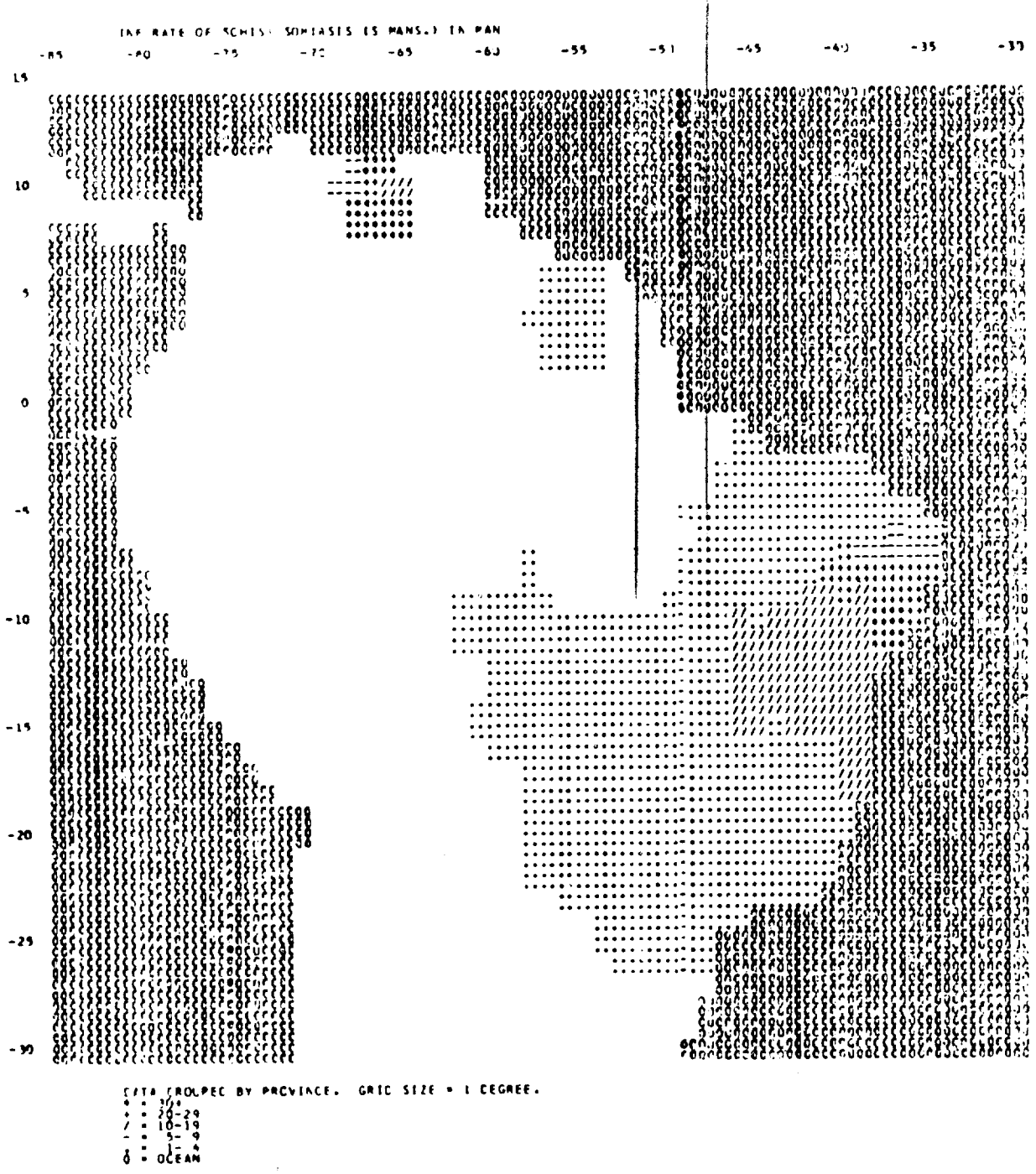


Figure 3-38 Shading-type map, produced on a line-printer by an IBM 7090, presenting standard set of South American schistosomiasis data (Fig. 3-1) plus oceanic points; this map demonstrates one method of adjusting computer output (by adding numerous zero-valued data points) to allow for natural features such as large bodies of water.

rule must be formulated to prevent contouring among data points so widely separated that linear interpolation becomes highly suspect. The alternatives to linear interpolation are quadratic, cubic, or higher-degree interpolation. Unfortunately, such methods, in addition to being more complex, usually require more points than do linear methods.

The only realistic solution to the problems posed by very sparse data points is to get additional *real* data. Any system, computer or otherwise, which gives credence to basically unreliable data does a great disservice, and we have made every effort in designing the MOD system to see that this will not happen. (For example, the MOD system incorporates a professional evaluation of the data, a CEN ((Computer Evaluation Number)), a means of identifying data in conflict, a NAR ((Narrative Output)) to supplement -- and point out limitations of -- the mapped data, etc.)

For production of contour-type maps, the manual algorithm described earlier (see Figs. 3-17 through 3-23) might be implemented on a computer. If this were achieved, the method would be at least as good as any program now existing. (This type of method has not yet been attempted so far as we know.) Existing methods of computer contouring almost always employ a grid (rectangular, triangular, hexagonal, etc.) for some aspect of the process of contour mapping. As a rule, a rectangular grid (usually square) is employed since this permits easy storage of data in an array. Even [illegible] which compute the surface statistically resort to gridding in order to display the results in map form. As in the production of shading-type maps, both the computer/plotter, and computer/line-printer configurations could produce feasible results.

Once the necessity of gridding data became apparent, the project team tried several ways of gridding, including circular, triangular and rectangular. One of our early experiments utilized the three-point-plane method suggested by Tobler (1964). Simply stated, it uses only the three closest data points surrounding a grid point, and fits a plane through

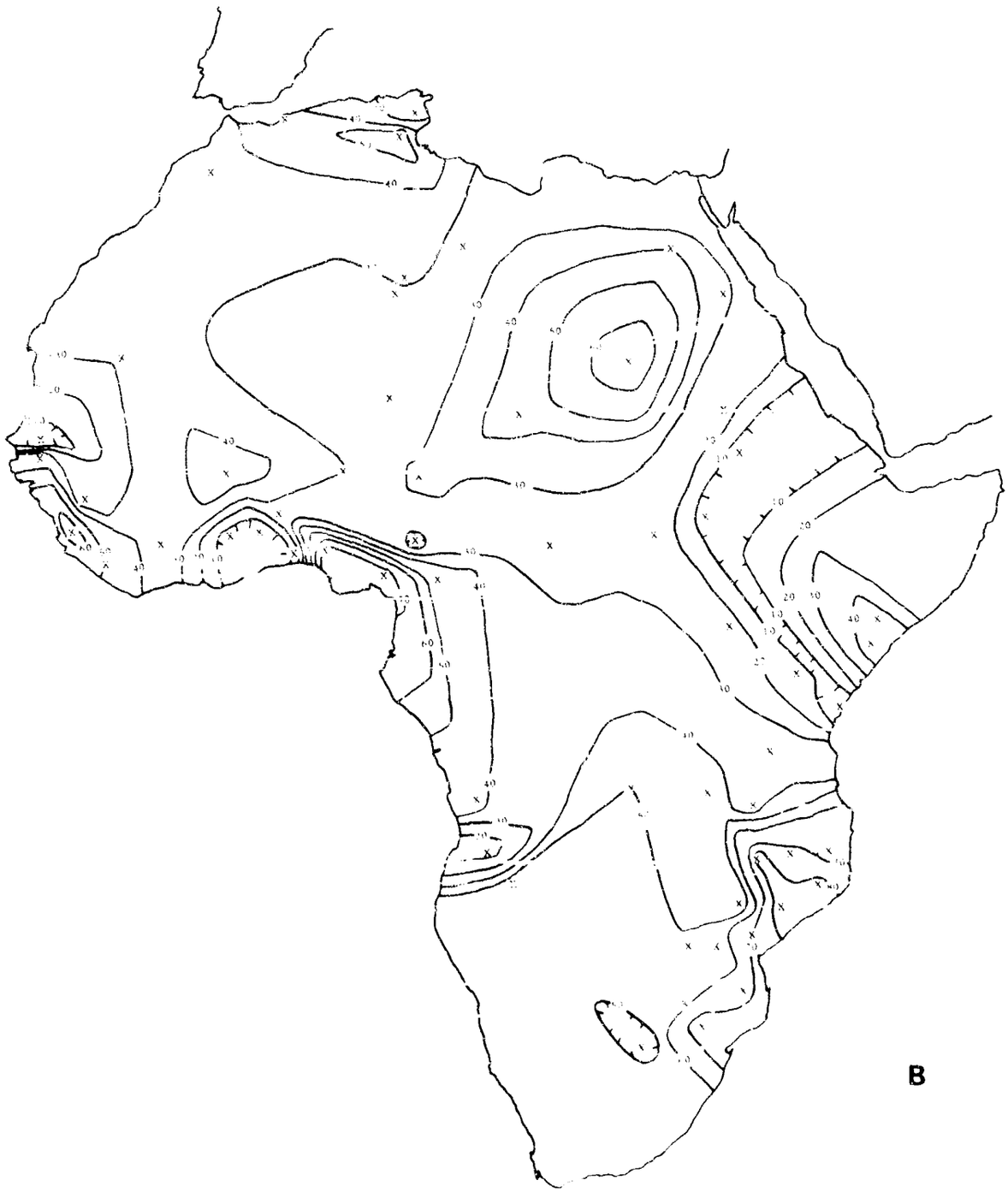


Figure 3-33 Manually drawn contour maps showing African schistosomiasis data: A, utilizing purely conventional cartographic methods; B, utilizing computer-like interpolation techniques.

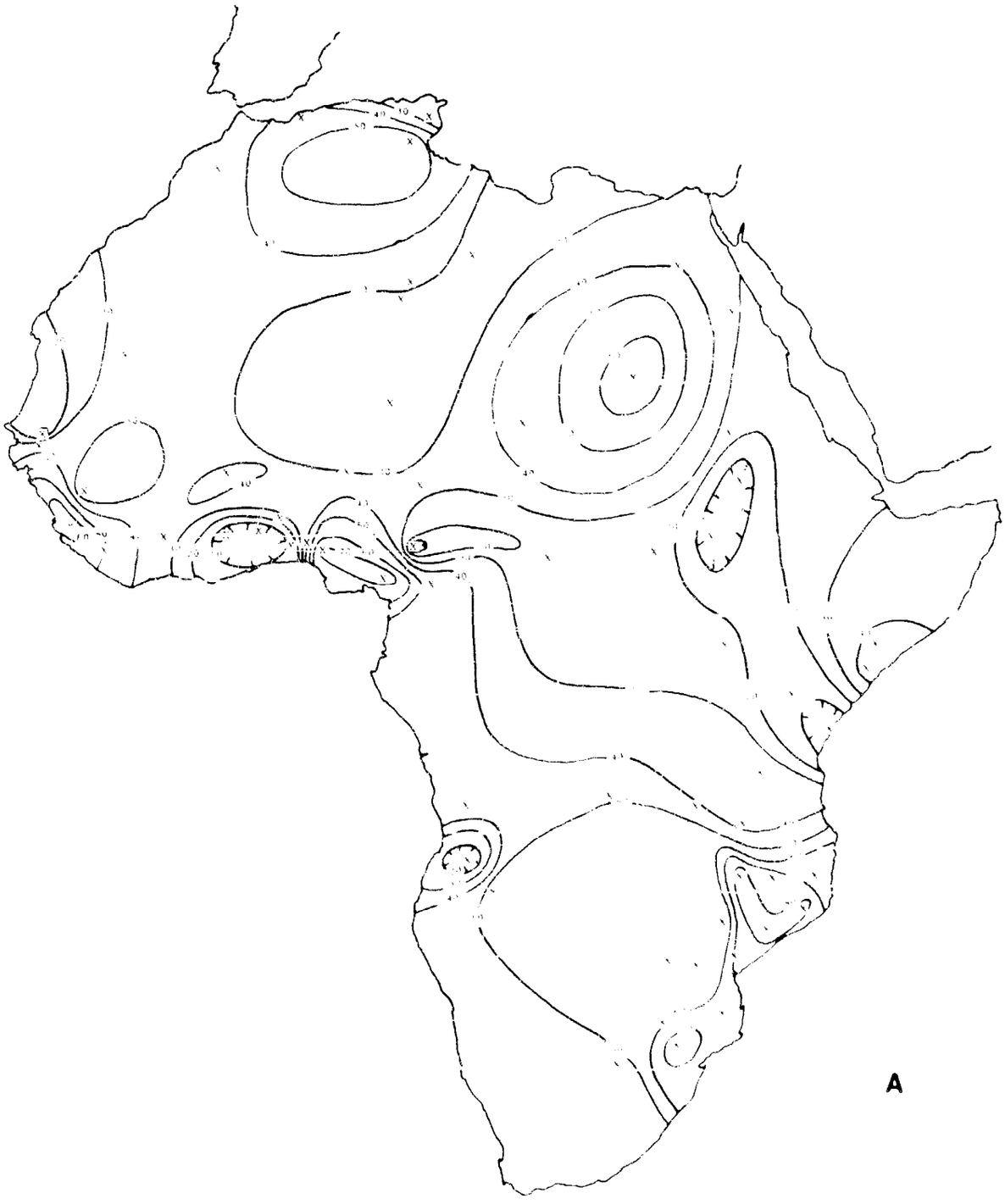


Figure 3-39A

them. Thus the grid point lies on this plane and its value may be computed. The method, as it applies to our system, is as follows:

(1) Compute the distance from the grid point to all observed points (which are assumed to be randomly distributed).

(2) Out of all these points, find the nearest three which surround the grid point in question.

(3) Fit a plane, $Z = AX + BY + C$, through the three points by solving the system of simultaneous linear equations necessary to fit a plane through points whose X, Y, and Z coordinates are known.

(4) Calculate the estimated value at the desired grid point by inserting its coordinates into the equation.

(5) Continue to the next grid point in the row; stop when all rows have been examined.

This method was performed manually to produce the map shown in Figure 3-40. Here, again, the standard set of (schistosomiasis) data were used to produce a contour-type map. All manipulations were carried out visually rather than calculated, but the concept can be demonstrated even by this rough approximation. A 1° grid was used for comparison with other methods.

An important point is that choice of grid size has a marked influence on the map which is [illegible] The coarser the grid, the more the map will indicate general trends; the finer the grid, the greater its detail. This is demonstrated in Figure 3-41, which uses the standard set of test data and a three-point-plane method of contouring, done manually. Note that if data points do not coincide with grid points initially, the net effect of gridding is to smooth out the data. See Figure 3-41 on the next page.

To evaluate existing computer contouring systems, the standard (schistosomiasis) data set was submitted to various existing computer contouring programs.

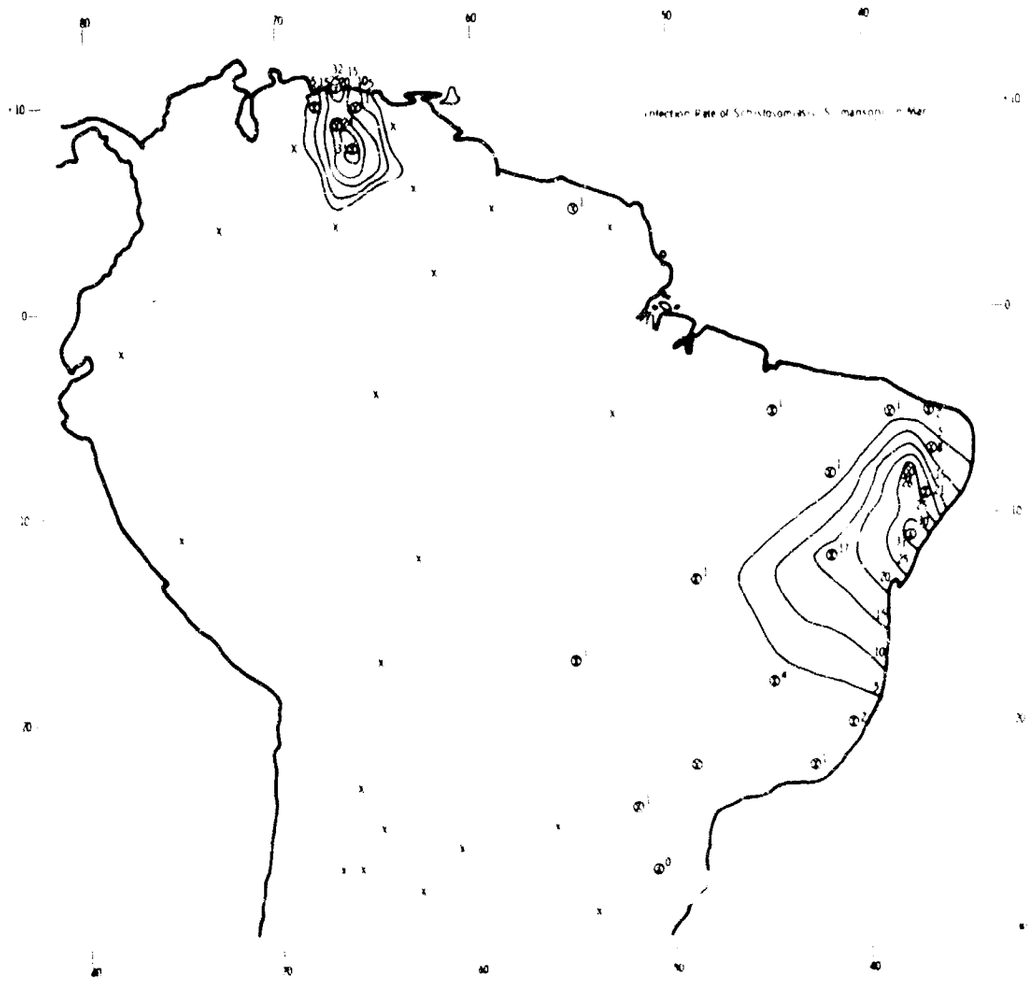


Figure 3-40 The standard set of South American schistosomiasis data (Fig. 3-1), mapped manually by a three-point-plane method utilizing a 1° grid.

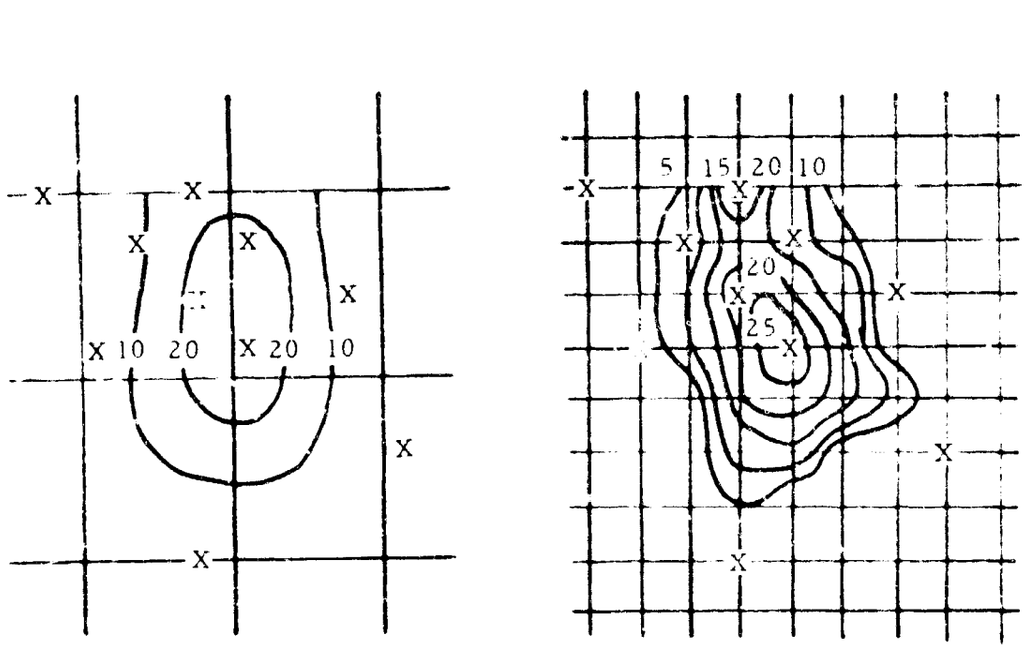


Figure 3-4l The standard set of (Venezuelan) schistosomiasis data (Fig. 3-1), illustrating effects of a coarse (A) versus a fine (B) grid during contouring operations.

The Control Data Corporation (CDC) offers at its data centers the most sophisticated contouring system generally available. Sufficient control data points must be present in order to calculate grid or mesh points on a rectangular grid. In addition, the original data must be randomly spaced. Thus accuracy is improved by adding points in areas of sparse distribution. The CDC program system performs essentially three tasks. The first task is to add points to the data to improve the accuracy of the computation performed in the second operation. Assuming the relationship between two adjacent points is linear, a linear interpolation is made to add points at levels which do not exist in the original data. An array is produced at the end of this operation, and is sorted and prepared for gridding.

The second task is to calculate grid or mesh point values for a rectangular grid. Grid point values are determined by finding the nearest known data points which surround a grid point and then calculating the grid

point value by an inverse distance function. Grid point values are computed only when data points surround an area, otherwise a "do not contour" value is attached to the grid point.

The third task is to contour automatically the three-dimensional data thus gridded. This data is expressed in X-Y coordinates with a Z value for contouring. The control values are stored within a matrix through which the program traces, interpolating to find the points through which the contour lines pass. The contouring is performed in strips of two adjacent rows of the matrix. Contouring is not performed where "do not contour" is indicated. The results of two parabolic interpolations are traced to compute the path of each contour line. As positions are calculated, plotter commands are stored in an internal array and output onto a plotter drive tape each time the array is filled. Optionally, the location of the data points, values, and grid lines may be plotted.

Our standard set of data was contoured by CDC, utilizing a variety of parameters. Figure 3-42 shows the data contoured using a 1° grid, for purposes of comparison with other maps (also because this was the level of accuracy to which the original data was given). Figure 3-43 shows the data contoured utilizing a much coarser grid -- approximately 2°. Here, only trends show. The map shown in Figure 3-42 presents some rather surprising results, and, as it stands, does not give an adequate picture of the disease situation (as compared with Figure 3-28, drawn manually from the same data by an experienced cartographer). The reason seems to lie in the manner in which the system manufactures "fill-in" data points. It can be seen that a relationship exists between each zero-value point and each positive-value point. Venezuela positive-data were combined with zero-value data through Colombia, Peru, Bolivia, and western Brazil to create a false impression. A similar false impression is given in the Colombia-Brazil border area. Another set of data was also used in investigating the CDC program. These dealt with rabies in the eastern United States and were produced from information supplied to us by the National Communicable Disease Center.

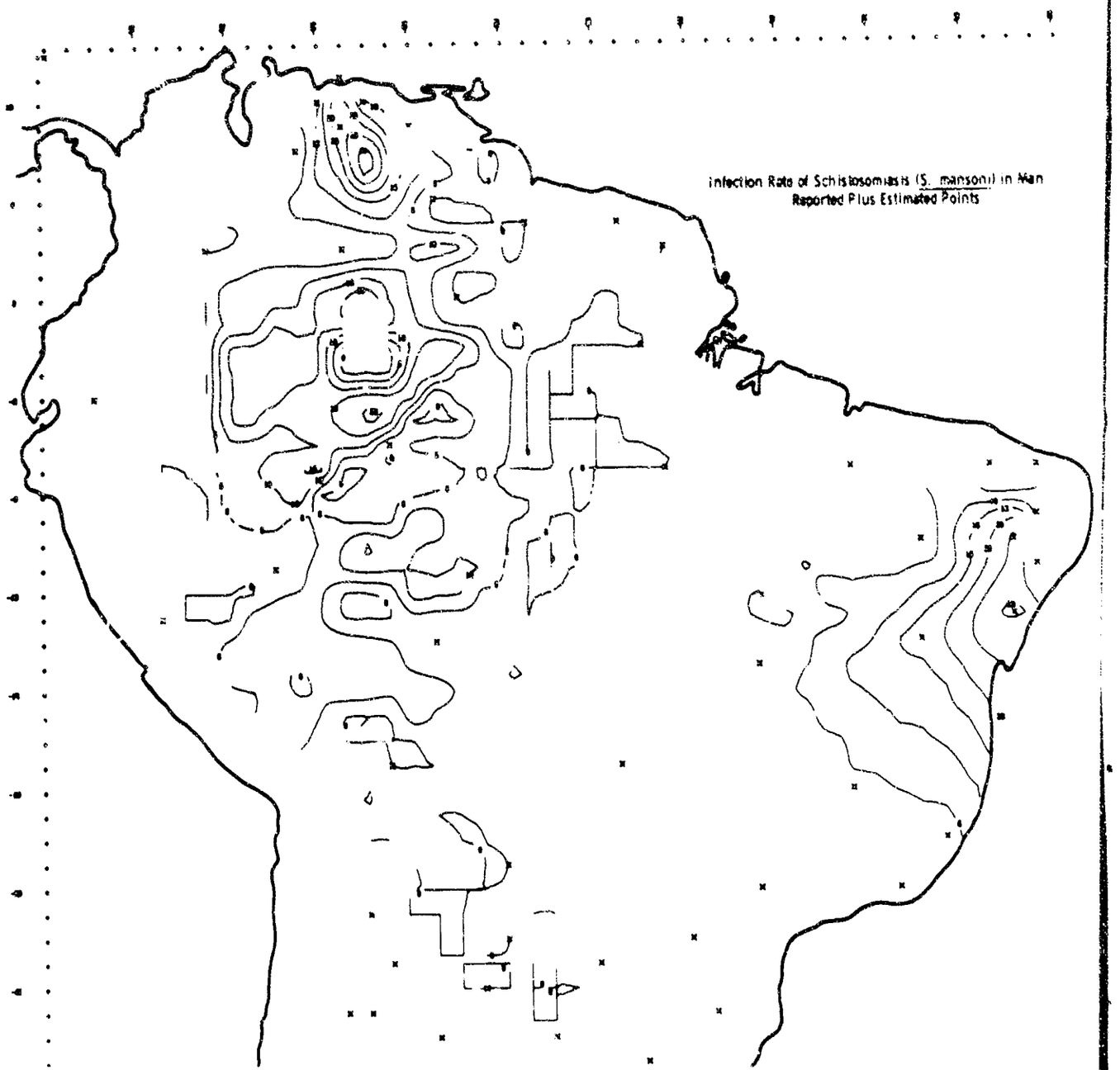


Figure 3-42 The standard set of schistosomiasis data (Fig. 3-1), contour-mapped by Control Data Corporation contouring system (using a CDC 3600 computer and an offline CalComp plotter), using a 1° grid (continent outline added manually).



Figure 3-43 The standard set of schistosomiasis data (Fig. 3-1) contoured by a Control Data Corporation program utilizing a 2° grid (continent outline added manually).

Figure 3-44 shows these data in the form of a map produced with the CDC program. Unfortunately, the MOD project was brought to a close before the implications of this experimental result could be explored.

The University of Michigan Geography Department (Tobler, 1967) has a line-printer contouring program which has been used successfully by them under a variety of circumstances, including situations with both large and small numbers of data points. This program utilizes a square grid for contouring. A simple smoothing technique, called the *moving average*, is applied to the data to make trends more apparent. The selection of grid size is determined automatically by the computer, which results in a coarse grid when there are few points and a fine grid when points are numerous -- based on the overall area and the total number of points. The grid size selected also determines the overall rectangle to be contoured (which can be somewhat larger than the original area). Grid point values are determined by one of two ways: (1), if the data point is less than an arbitrary distance away from the nearest grid point, its value is used for the grid point or (2), if the data point is not sufficiently close to the grid point, it is determined by a weighted average between the closest point and the six closest points (the closest point is included in the "six closest points").

Our standard set of test data was contoured by their program, and the result interpreted to produce the map shown in Figure 3-45. Figure 3-46 shows the computer-generated output map. This map shows the effect of a large grid (i.e., it presents trends only); the grid size, approximately 8°, was selected automatically on the basis of total number of points and the area covered.

Most of the existing computer programs for mapping are designed to use a large number of data points. When only a small number of points is available, trends may be the only meaningful result which can be obtained. In cases of this sort, mathematical techniques can sometimes be applied to all the data points together, and will give a better indication of trends



Figure 3-44 Unpublished rabies data from National Communicable Disease Center; A, contoured by a Control Data Corporation system (equirectangular projection); B, contoured manually (Albers equal-area projection) -- courtesy of V.T.Carofalo. As was the case with the schistosomiasis maps, the computer-drawn map indicates only broad trends in the data and is not precisely similar to the manually drawn map.



Figure 3-44B

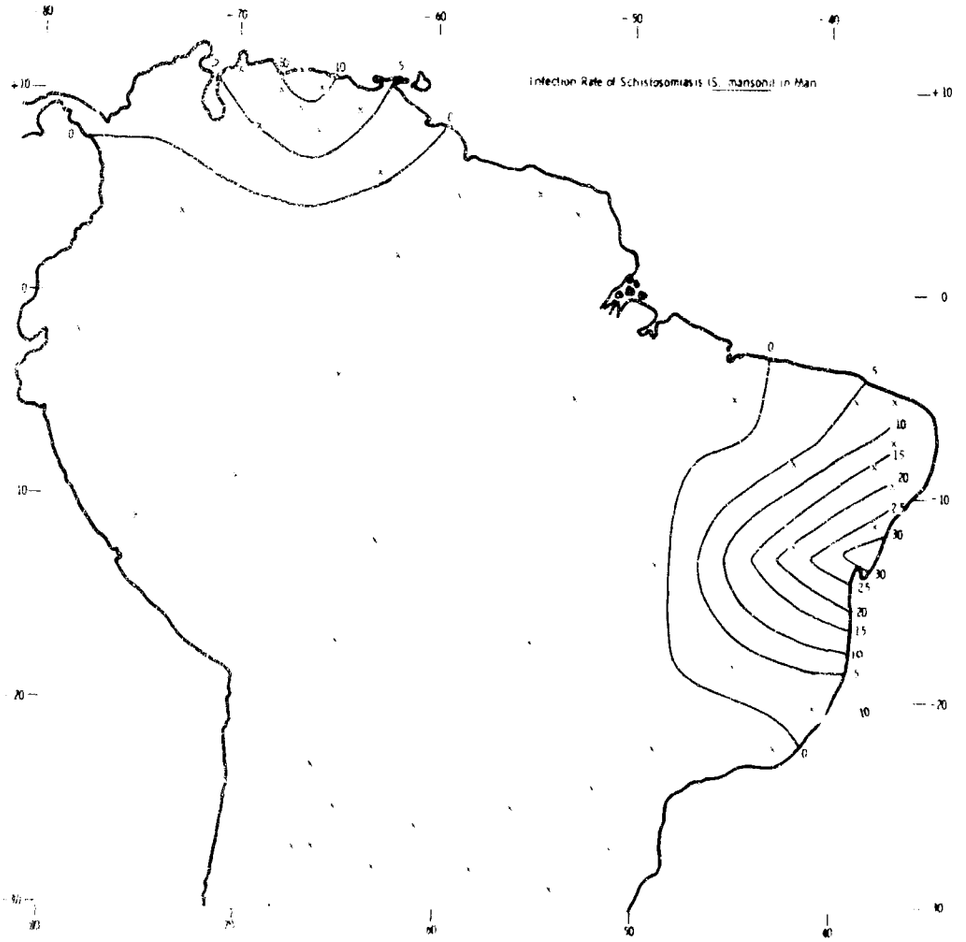


Figure 3-45 The standard set of South American schistosomiasis data (Fig. 3-1) machine-mapped using University of Michigan contouring program. Contours in this figure were manually traced from the output shown in figure 3-46, on next page.

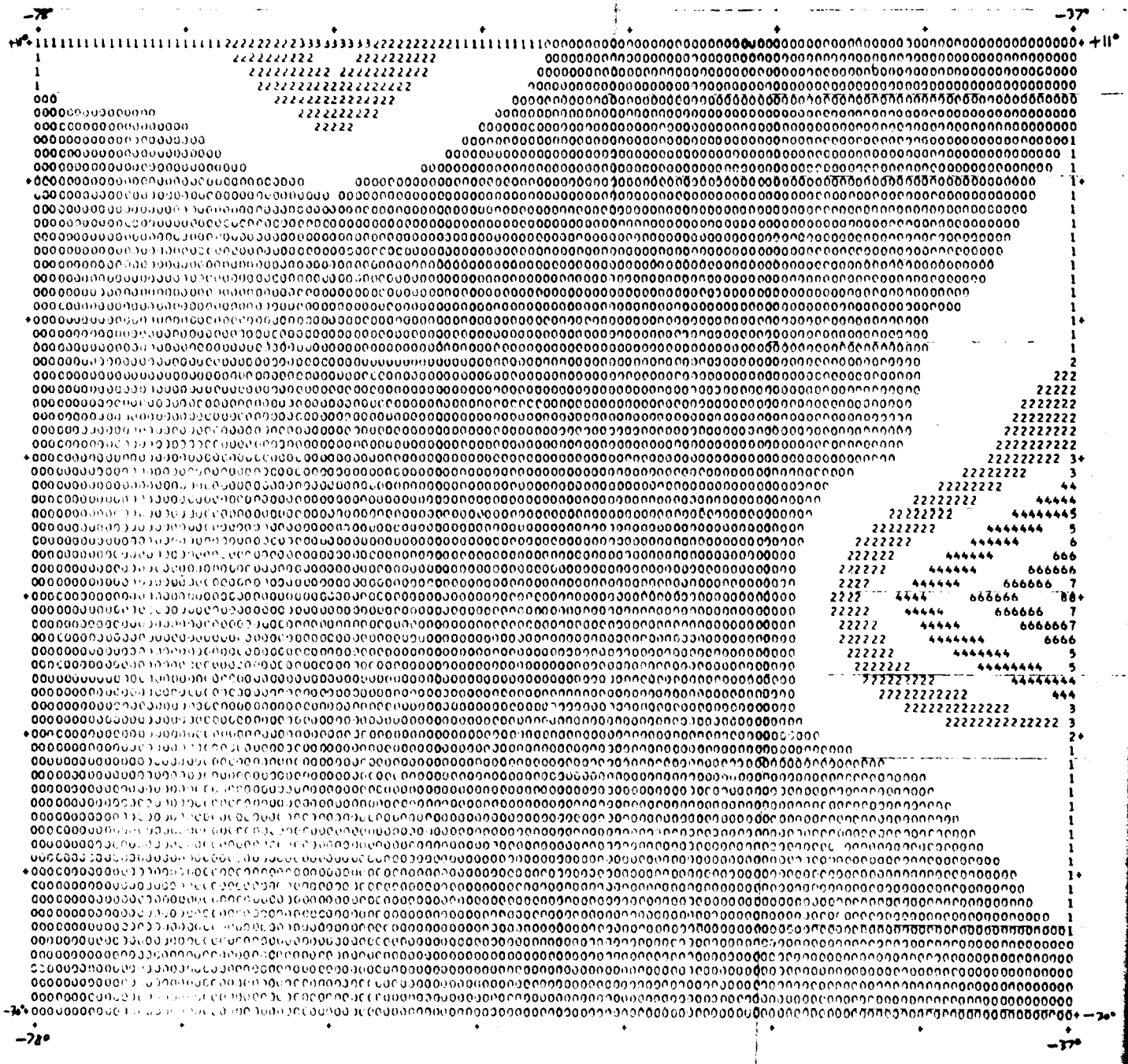


Figure 3-46 The computer (IBM 7094)/line-printer produced map from which were taken the contour lines shown in Figure 3-45.

than standard mapping techniques. Such three dimensional data can often be summarized and interpreted by reduction to a simple geometric shape. A geometric surface can be fitted to the data by an extension of the least-squares curve-fitting procedure used in curvilinear correlation and regression analysis. The resulting fitted surface is an approximation or "trend" of the data. A sixth-degree trend surface requires a minimum of 28 data points (28 points would give an exact fit since residuals would be zero), and we explored this approach.

A computer program developed at the University of Kansas (O'Leary, Lippert, and Spitz, 1966) for use in geological applications, was used to calculate a sixth-degree polynomial surface for the standard set of (schistosomiasis) data utilized in these map studies. The program was also used to contour the results on a line-printer. This program system first calculates a coefficient matrix and column vector, then solves the matrices, ordering before each elimination. It then calculates and prints trend surface Z values (for degree one through degree six), residuals, error measures, and equations of surfaces. Next, the trend surfaces requested are calculated and printed as a contour map using alternating bands of symbols and blanks to identify contour lines and the regions between these lines. If desired, Z values (Fig. 3-10) and residuals are plotted as a dot-type map on the printer. Figure 3-47 shows the interpreted computer output in a form similar to the other results, and Figure 3-48 shows the computer-generated output map.

This method should normally produce acceptable trends, but, in this example (Fig. 3-47), it fails in the Peru-Colombia-western Brazil area, presumably due to the lack of control points in that area. On this set of test data the University of Michigan program gave a somewhat better result than the sixth-degree polynomial method. In general, the weighted moving average gives results comparable with those obtained by the fitting of local polynomials, and the result is computationally much simpler. (A different polynomial fitted to these data points would have given a different, though not necessarily more acceptable, trend surface.)

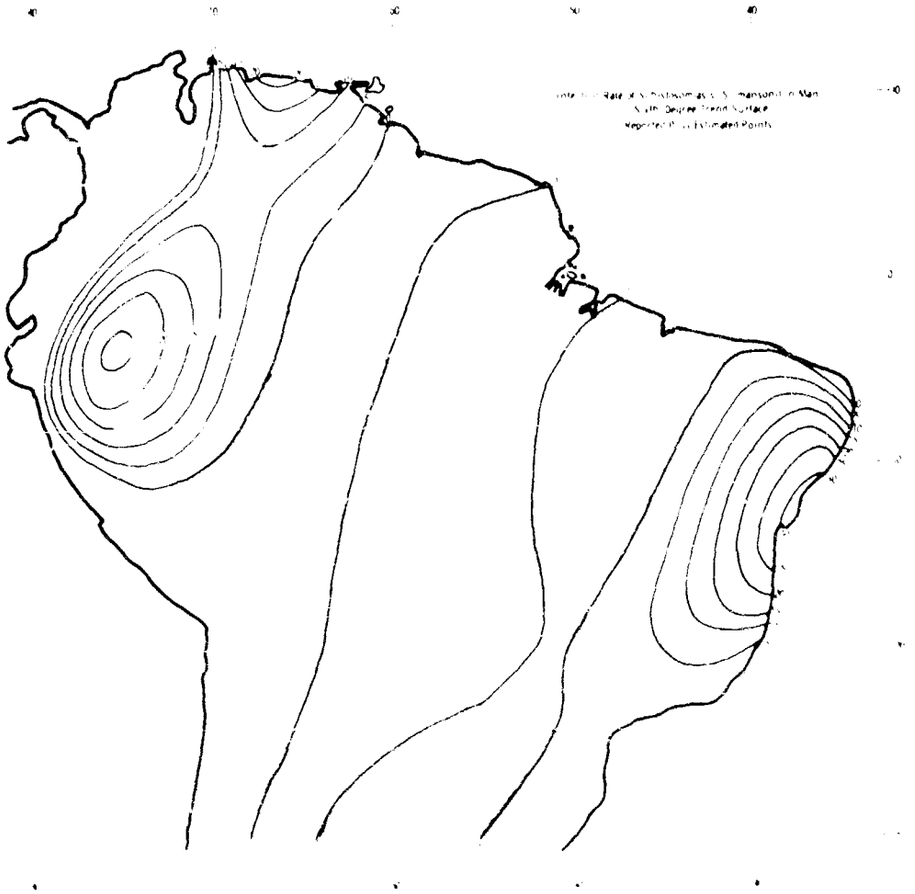


Figure 3-47 The standard set of South American schistosomiasis data (Fig. 3-1) contour-mapped as a sixth-degree trend surface by the Kansas Geological Survey trend-surface program. Contour lines were traced manually from the output map shown in Figure 3-48.



Figure 3-48 Computer (IBM 7090)/line-printer produced map from which contour lines of Figure 3-47 were drawn.

Other methods have been studied to a limited extent, but, in general, have not been found to be applicable. For example, a double-Fourier series program exists, but requires more points for accuracy than will usually be available; vector and factor analyses are, as a rule, not suitable.

Throughout these experiments several techniques were evaluated, testing alternate methods,etc., seeking ways to improve output. For example, we tried using the same value for each grid point falling within a single political unit. Compare the base map shown in Figure 3-36A with Figure 3-49; figure 3-49A shows the result if the data value for each province are entered at each grid point within the province before contouring. Figure 3-49B is a map based on the same data, but this time the data points were located at the center of each province. In general, spreading out the data is more useful for shading, whereas averaging the "center point" data is more useful for contouring. Both methods could probably be applied to both types of mapping, within certain limits. For example, when the data are spread out over a grid, they appear to be more suitable for shading as the grid is made smaller -- and more acceptable for contouring as either the grid or the contour interval is made larger. Figure 3-49A would be improved if either a coarser grid or a larger contour interval were used. Alternatively, center-point data are better for contouring when the points are fairly randomly spread over the area to be contoured -- and better for shading when the area boundaries are not known. Figure 3-49B does not have an adequate number of points in the lower half (based on the grid size used), and the contour lines wander accordingly. When the extent of an area described by a data point is unknown, the data value could be carried half-way to the adjacent data point for shading (as shown by Fig. 3-36D). This method is described by Harvard University in their SYMAP system (Fisher et al, 1967).

When area boundaries are known, it is preferable to spread out the data over a grid for shading by computer because the resulting shaded map is reduced to a dot-type map with a very large number of dot positions, and each dot position can be filled with the appropriate symbol. In Figure 3-37I,

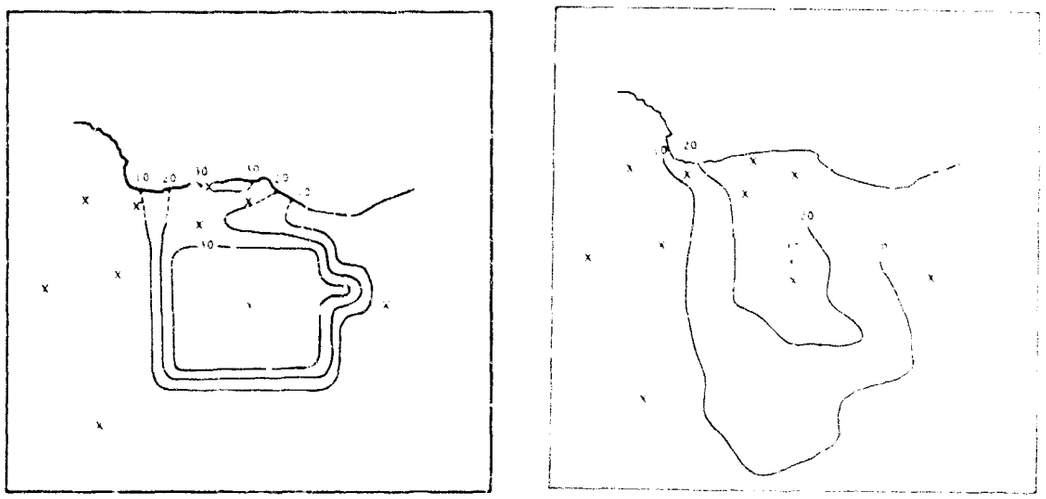


Figure 3-49 The standard set of (Venezuelan) schistosomiasis data (Fig. 3-1) contour mapped (compare with Fig. 3-36A): A, with data for each province spread out over a $1/2^{\circ}$ grid; B, with data for each province grouped at its (areal) center.

```
??????????14??????????????????????????????????09????????????
????????????????????21????????????????15????????????????05??
??????????????????????????????????????????????????????????????
????????????????????32????????????????????????????????????????
??25????????????????????????????????35??????????????????????????
????????????????????????????????39????????????????????????????
??????????40????????????????????????????????????????????11??
????????????????????51??????49????????????????????32??19??????
????????????????????????55??????5968??????39????????????????
38????????55??73??61????????????????????????????????????????
??????????????????????????????????????42????????????????
????????????????54????????????????????????????????????????
????????????????????????????????????????????????17???????
????????41????????????????????39??????????????????????????
????????????????????38??????????????????????30????????????
??????37????????????????????????????????????????????????
????????????????????????????????????????????4?????????
????????????????????????????????????????????????????08
????????????????????????????????33????????????????16??????
????????35????????????????????????????????????????????????
??????????????????????????????????????????????????12????
```

Figure 3-50 A line-printer-drawn map showing (by numbers) the known data points of Figure 3-18, and (by ?'s) the unknown data/grid points in between.

the simulated shading would have been done by a plotter, a line in one box connecting with another line in the next box (actually, the entire line would be drawn in one stroke, as the computer would test for the end).

Since data are gridded for contouring, and the grid size determines the form of the output, we considered techniques for varying grid size on the same map, but with little success. As an alternative, we produced separate maps (with different grid sizes) for various parts of the overall area and joined these manually. (Varying the grid sizes requires manual adjustment to line up the contours when the maps are fit together.) Lack of time prevented further investigation of variable gridding techniques, and several ideas remain to be explored. For example, a grid file could be set up to include different-sized grids in different regions, based on medical workers' empirical recommendations, alternatively, grid size could be varied, depending on the number of data points and their average spacing.

In another experiment known data points were represented by appropriate symbols and all the unknown grid points were also represented. The resulting map (Fig. 3-50) is very difficult to understand although it is derived from exactly the same simple data as depicted in Figure 3-18. From this it appears that, as a rule, unknown points should not be indicated, as such -- and this conclusion is in line with standard cartographic procedure.

Information as to unknown data points can be implied by presenting dot-type symbols on a shaded or contour map. From the dot-type symbols it will be quite evident where known data were used and where interpolated values were added. With this information, the knowledgeable user can challenge questionable areas (and, perhaps, set about obtaining data for these areas).

In another experiment we compared the results of a general or trend map (Figs. 3-28, 3-40, 3-43, 3-45, and 3-47) with a more detailed map (Fig. 3-51). The data used in the previous studies were plotted as accurately as possible without grouping and averaging it by province. These data were

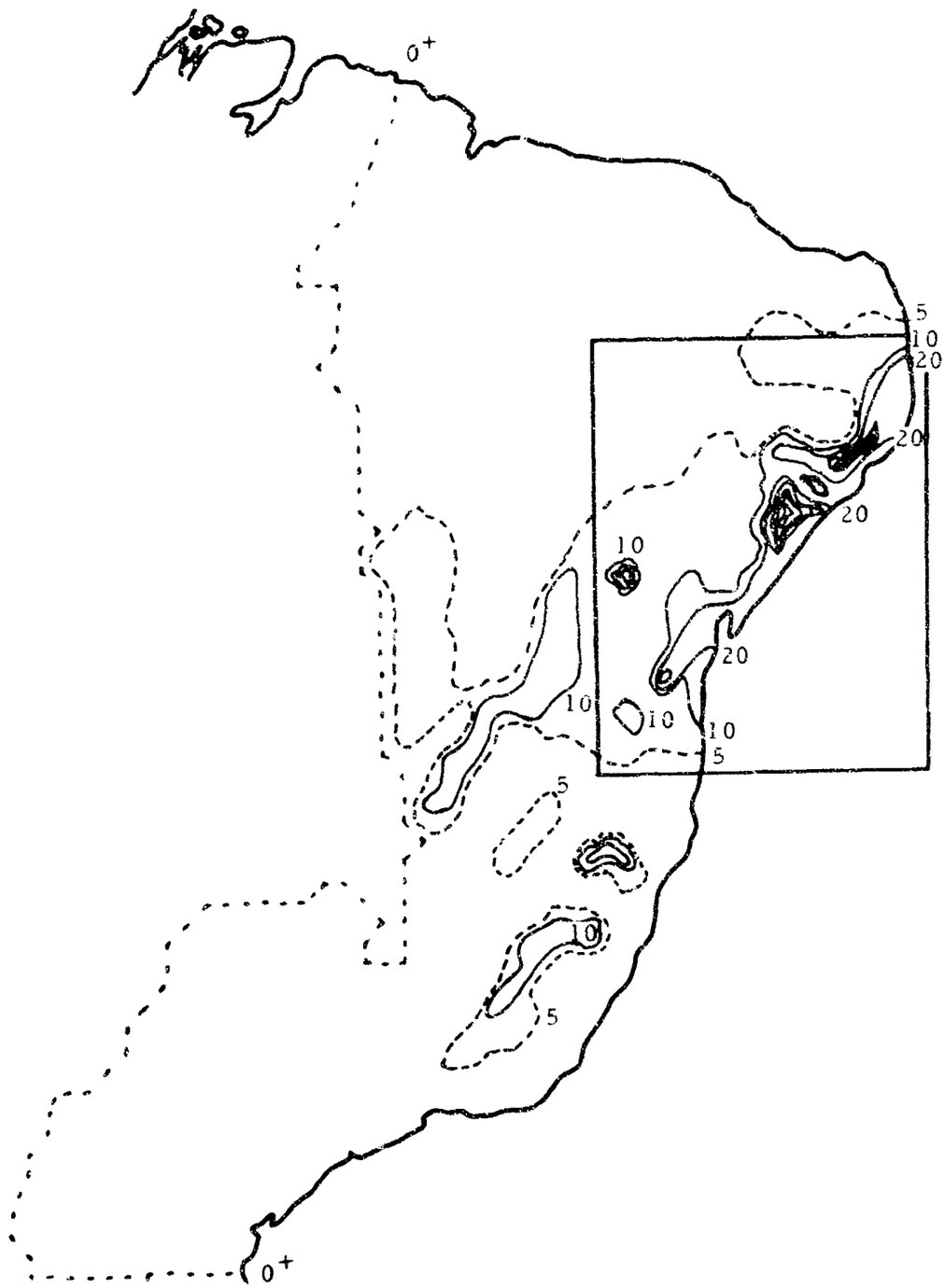


Figure 3-5l, A and B Schistosomiasis data for eastern Brazil (from Malek in May, 1961, p. 305-313) contoured by spreading data over entire extent of each reporting area rather than by taking data at "center-of-gravity" of each reporting area or by grouping all data into province-sized reporting areas -- A, overall region. (3-51B is on page 3 - 93)

then gridded and contoured manually. Needless to say, trends are not as clearly distinguishable on this overall map (Fig. 3-51A). But if we look at Figure 3-51B, which is an enlargement of one section of 3-51A, it appears that this technique of grouping data may be more useful in small-area studies. MOD experimental maps produced in this manner compare surprisingly well with published maps presenting comparable information (Fig. 3-52).

One factor that pertains to contouring and which places some limitations on the output is that data extremes and data planes cannot be contoured. Note the zero contour line in Figure 3-42 for example. The computer assumes this line is ever-present anywhere in an area of zero data values, and is quite unsuccessful in contouring it. This is because zero is the lowest value on the map and, in order to contour, it is necessary to have values both lower and higher than the contour line to be drawn. It might not be apparent, but a plane composed of several values which should fall on a contour line, likewise, cannot be reconciled by a computer. This is exemplified by Figure 3-53, which represents a section of a map produced from the standard MOD test data by the Naval Oceanographic Office (Osborn, 1967). The grid size used was about 1 1/2°, and is quite apparent on the map. The 1% contour line, labelled 101% on the map, appears as a plane (identified by the arrow) on which the computer traces between all grid points.

A common solution could be applied to both these problems. Before contouring, all grid point values could be offset downward an identical minute amount to insure that no value falls on a contour line. This would make all values slightly lower than reported, and would thereby permit construction of the zero contour and all other contours, as any planes would fall between contour lines.

As a final study of gridding, see Figure 3-54, where we have reproduced figures from a gridding study performed by the University of Kansas (Preston & Harbaugh, 1965); Figure 3-54A shows a very detailed topographic

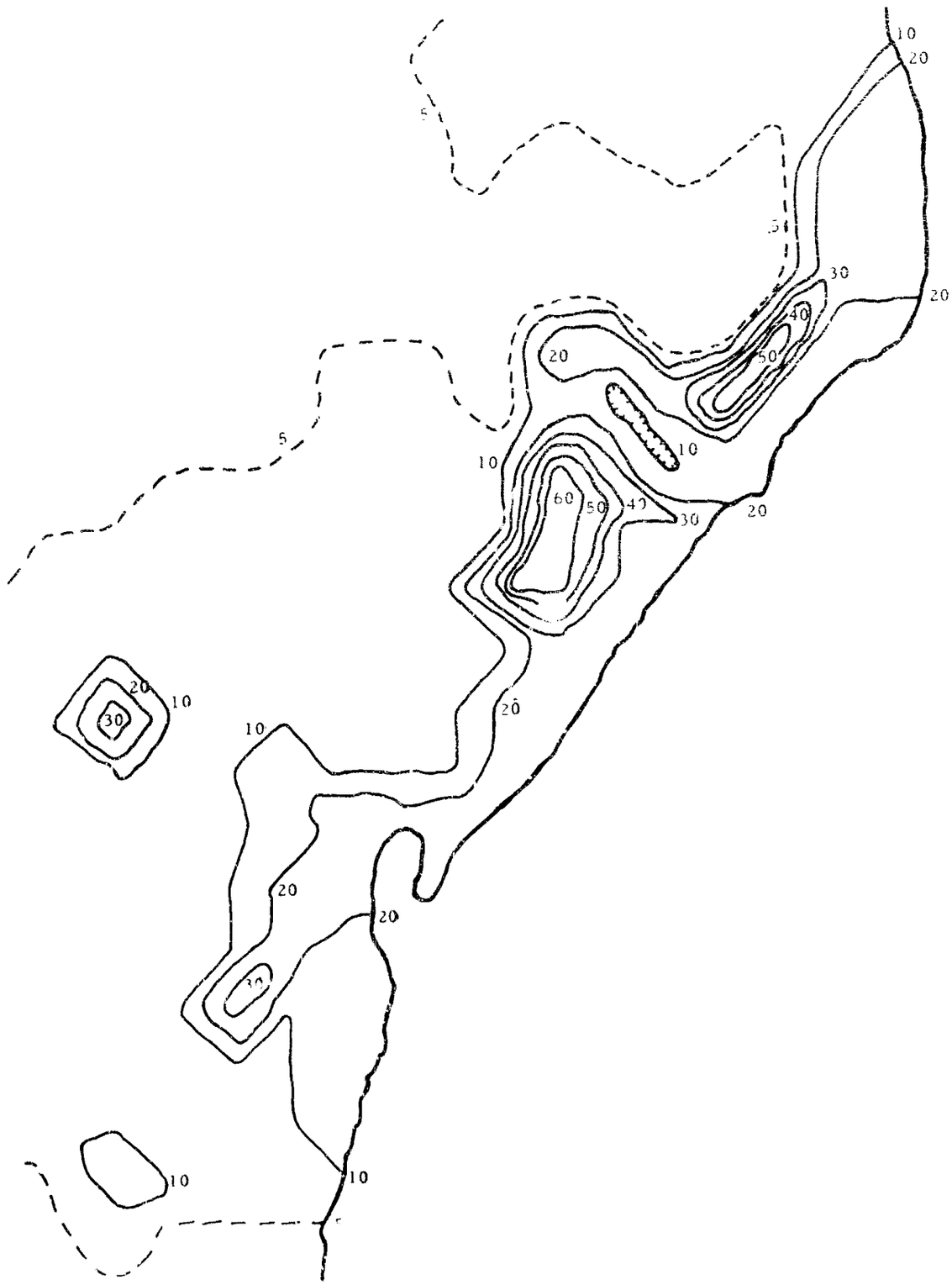


Figure 3-5lB The outlined region of -A enlarged.

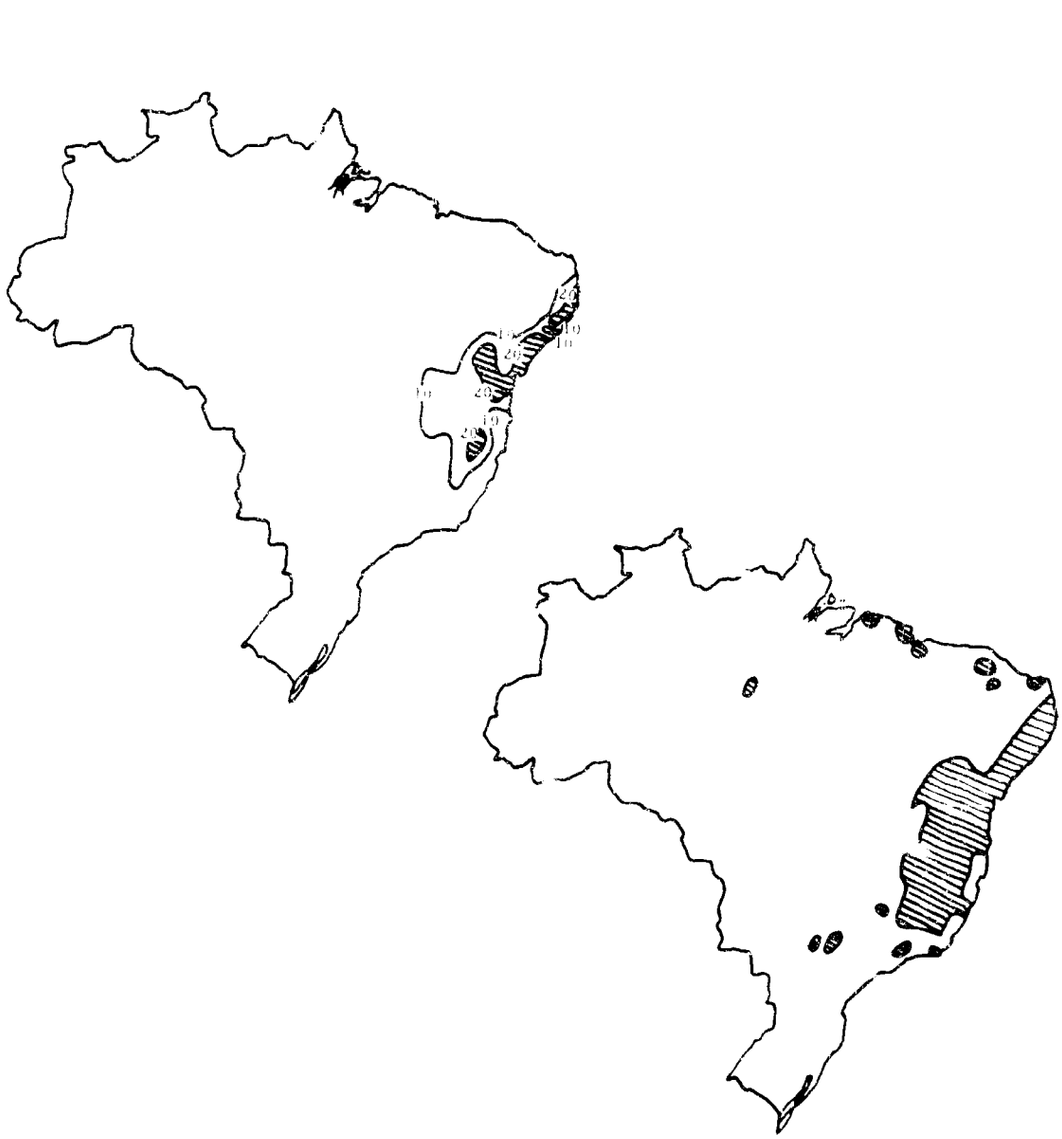


Figure 3-62 Schistosomiasis data for eastern Brazil; A (upper), MOD-produced map (from data of Malek in May, 1961, p. 305-313); -B (lower), publisned map showing comparable data (shaded areas indicate 30% + prevalence) -- from Pathology in Brazil, Past and Present. International Pathology, vol. 8, p. 8, 1967 by Domingos de Paola, used with permission.

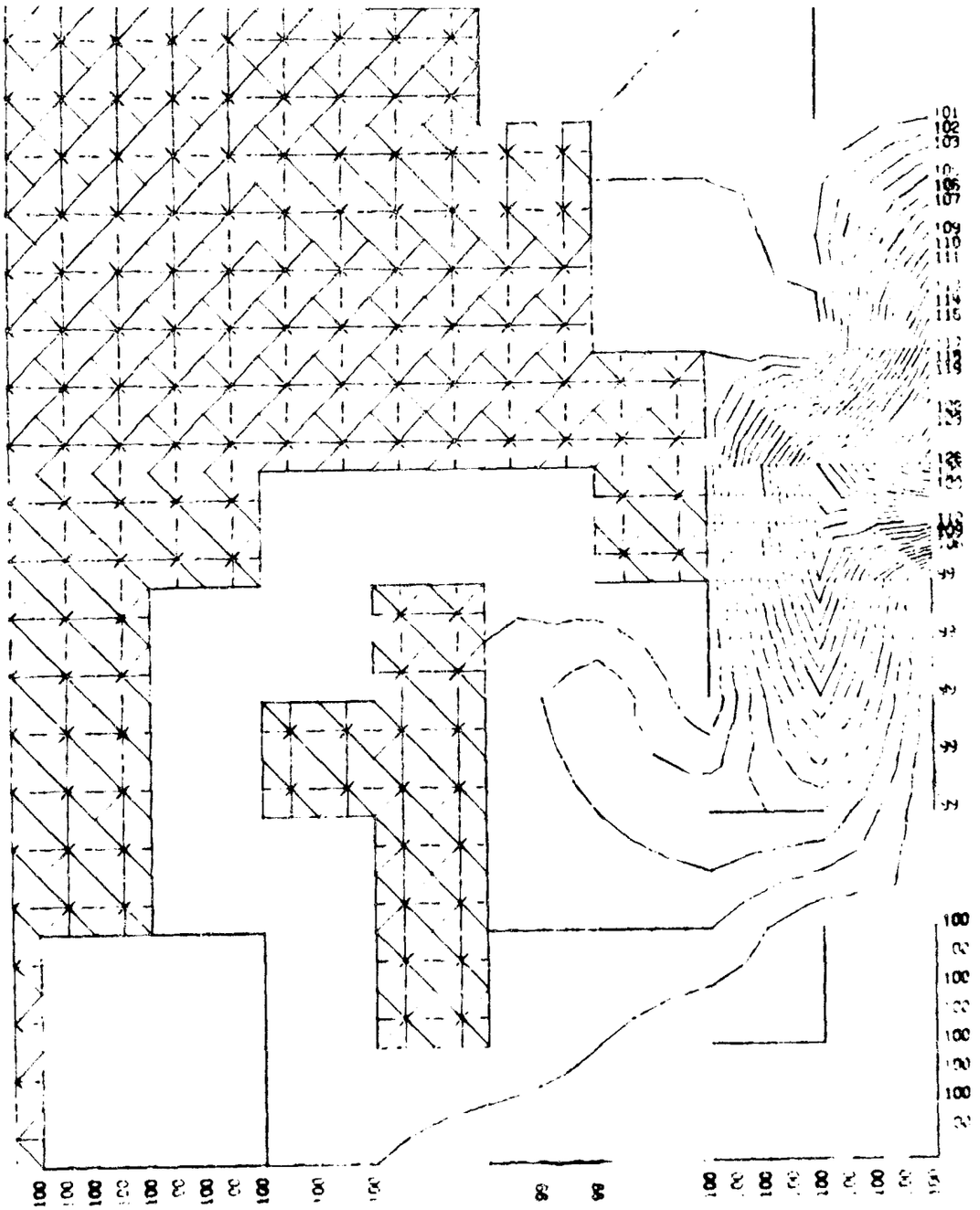


Figure 3-53 The standard set of schistosomiasis test data (Fig. 3-1) for eastern Brazil contour-mapped by the Naval Oceanographic Office program. The arrow points to an area in which (because contoured surface is a flat, horizontal plane with a value of 1% infection rate) the computer traced between all grid points of equal value.

map. Data values were taken from this map at regular intervals (which represents the construction of the grid) and recontoured manually (Fig. 3-54B) to show the best representation possible from this size grid. Note that the resulting gridded map compares favorably (but not perfectly) with the original topographic map. It would be interesting and valuable (and ultimately necessary) to take a map of a single environmental factor, digitize it (by gridding), and reconstruct the map using the MOD system -- and this will be one of the crucial tests to judge whether or not the computerized mapping system is ready for use.

In this consideration of map construction, we have concentrated on problems involving data points -- data grouping and interpolation between "real" data (groups), particularly in connection with computer program/plotter operations -- and this emphasis was appropriate. We should not conclude our discussion, however, without at least a brief consideration of accuracy requirements in computer/plotter operations. In a recent presentation (1967), M.A. Richardson and J.S. Rollett discussed this problem (and other important ones, e.g., quantities of point and line information required for an effective data bank, etc.). They described some of the critical situations (relating to accuracy), including several of particular interest to our contouring activities:

- Crossing of lines intended to run closely side by side.
- Kinks on lines and failure of loops to close.
- Failure to place a point symbol symmetrically on a line.

It was their conclusion that "... the machines used should be such that not more than 1% of errors will lead to mismatches of features which exceed 0.006", and ... this implies an accuracy of positioning a given feature of plus or minus 0.003"." It should be noted that these statements related to "a computer compatible system for automatic cartography" capable of producing complete (conventional-type) high quality maps. Furthermore, their concern was not so much in whether or not the machine produced visible

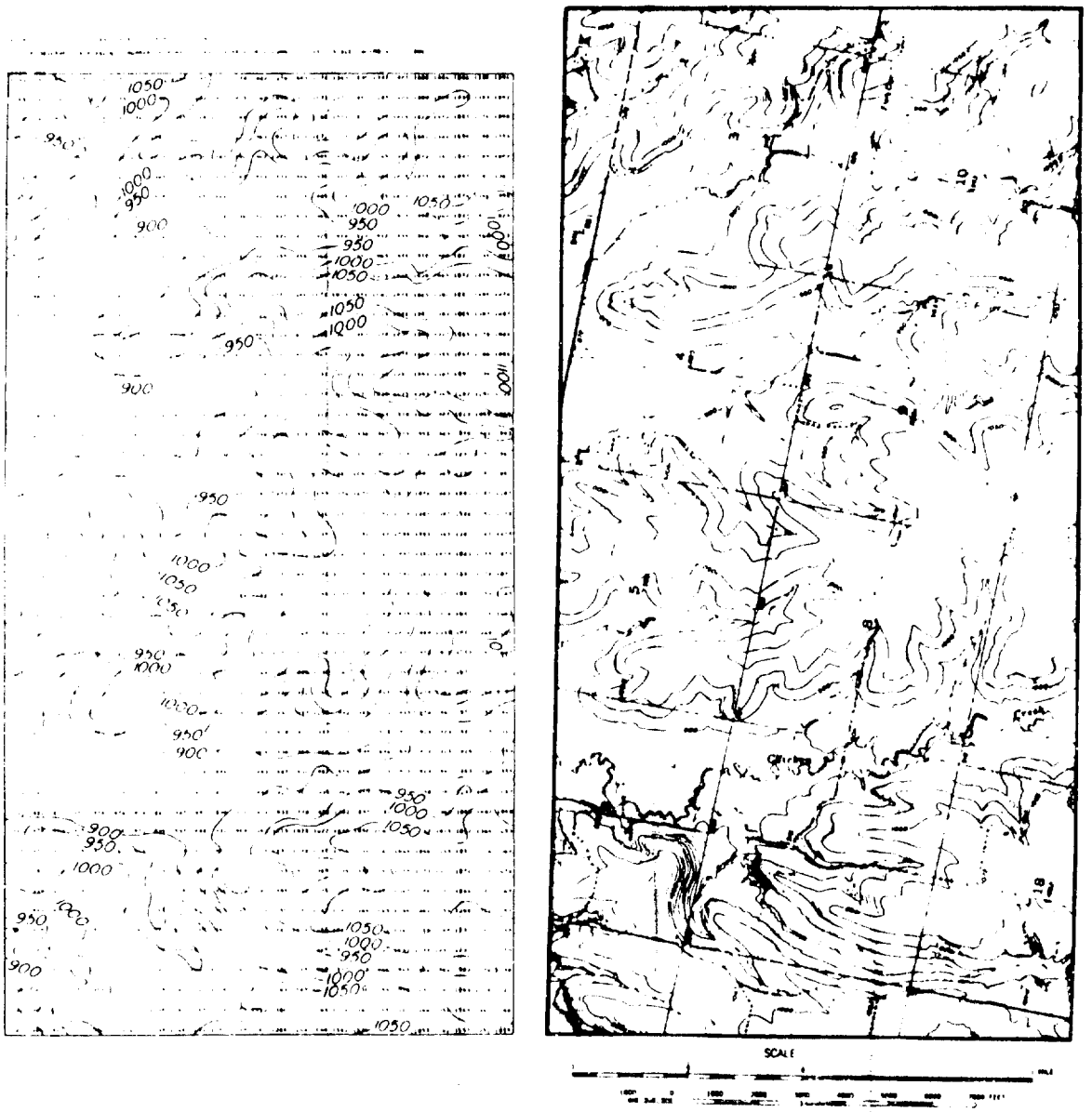


Figure 3-54 A, Topographic (contour) map of the northern part of the Lone Star Quadrangle, Kansas, showing an area 4.5 x 2.25 miles; -B (left), data point values taken from A (right) at intervals of about 0.1 mile according to a rectangular grid pattern, printed out on line-printer and re-contoured manually for comparison with A, the original map. Although much detail is lost by gridding data, the broad trends are still quite evident -- Preston & Harbaugh, 1965, p. 64-65; courtesy of State Geological Survey of Kansas.

defects, but, rather, in how often these were likely to occur and how much (human cartographer) effort would be required to put things right. Their requirement of ± 0.003" accuracy was chosen because it is within the limits of an economic load of hand corrections.

The MOD system, of course, has a different objective; it is concerned with production of maps which, as a rule, will present limited kinds of data on one sheet to be used in conjunction with base maps. As we have discussed before, the medical-environmental data are of more abstract nature than roads, mountain peaks, ponds, territorial boundaries, etc., and their positions must be calculated. Since, at best, these calculations represent (close) approximations, the degree of accuracy necessary in MOD system maps is not nearly so great as that required by Richardson and Rollett. Although we have not explored accuracy requirements in depth, it is our opinion that several plotters, currently available, will satisfy requirements for the MOD system -- at least for the present. The weakest link in the long chain which contributes to inaccuracies is the link that represents the raw data. Next in order of importance is the link which represents the method of interpolation between real data-points.

3.3 BLOCK DIAGRAMS

Block diagrams are an offshoot of maps in that they attempt to give a perspective view of the form of the land or of a statistical surface by presenting it obliquely; recall that, upon viewing a map, the orientation is perpendicular to the mapped surface. The term "block diagram" is well-established in the geologic and geographic literature; "perspective drawing" is also used (but often with a broader meaning). Block diagrams are often called three-dimensional maps, but the only true three-dimensional maps are "relief maps". Histograms may also be prepared in perspective, attempting to depict three dimensions. These are essentially block diagrams of a three-dimensional step-function (which could also be represented by shading-type maps, with the data based on area boundaries).

Figures 3-55 and 3-56 show block diagrams produced to display non-disease data. The standard set of test (schistosomiasis) data (Fig. 3-1), used by the MOD project to produce various types of maps, is also shown here as block diagrams, first, as drawn manually (Fig. 3-57), then, as drawn by a computer/plotter at the University of Michigan (Fig. 3-58).

The program which produced this last-mentioned block diagram (Fig. 3-58) is the only such program that was investigated in the time available to us. Processing of the data for the block diagrams was done by an IBM 7094 computer at the University of Michigan; plotting of the diagrams was performed by an off-line CalComp plotter. This program successfully represented the standard set of schistosomiasis data in a vivid block-diagram format; however, further investigation will be necessary to test adequately these techniques.

3.4 GRAPHS

We have touched upon the usefulness of graphs, in general, and their possible application to the MOD system (3.1.2). Let us now explore a little deeper the idea expressed there of a "family of curves", each curve considering three major disease-environmental factors -- two of these variable, the third held at a particular (fixed) level. An approach of this sort could be of great value to the research epidemiologist or others interested in exploring precise relationships among various causal factors of a particular disease. The principal deterrent to this approach is the "depth" and extent (geographically) of the data that are required. Among other things, it appears that these restrictions would almost require that the area under consideration be a relatively small one. We consider the following hypothetical disease-environmental situation to illustrate the application and potential usefulness of the method. The two variables considered in the graph are: (1) prevalence of human leptospirosis, and (2) salinity of surface water; the fixed factor is the pH of surface water.

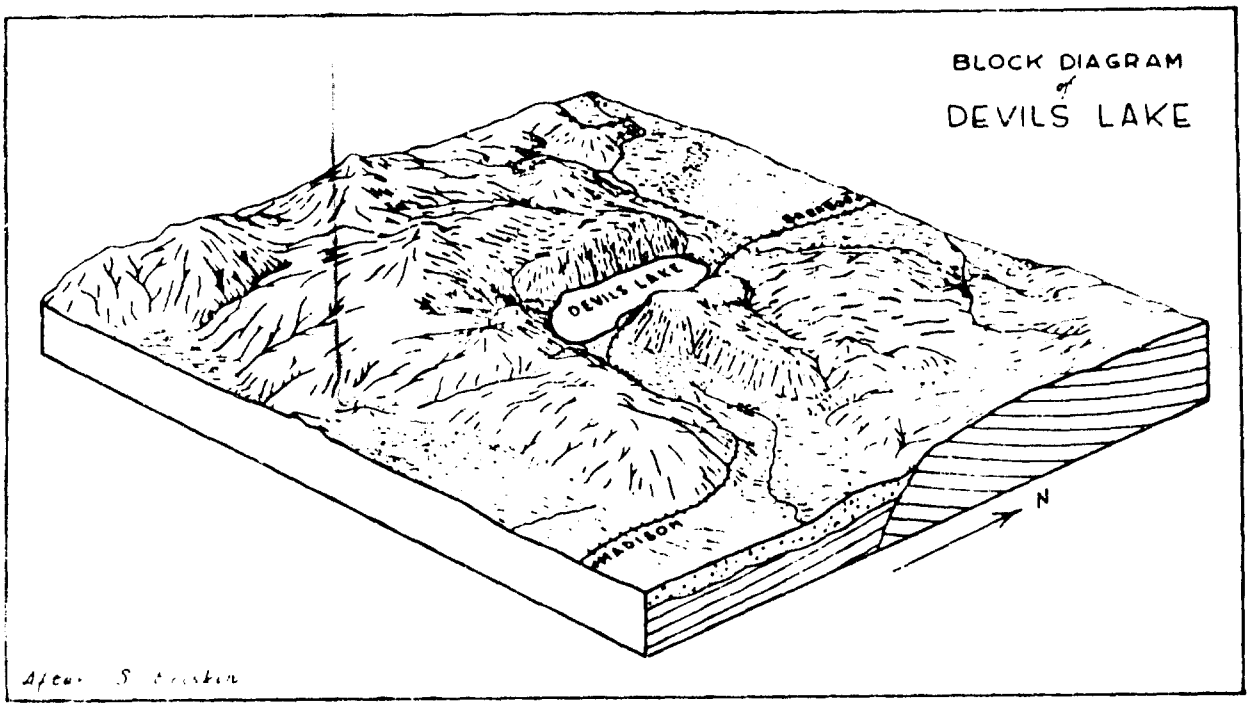


Figure 3-55 Published, manually drawn block diagram portraying the form of the land's surface near Madison, Wisconsin.

from Elements of Cartography, 2nd ed., by Robinson, A. H., 1960, published by John Wiley and Sons, Inc., New York and reproduced with permission.

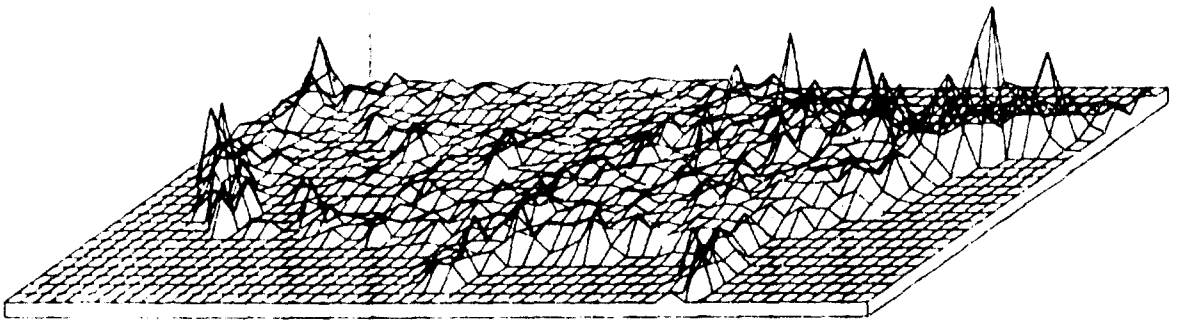


Figure 3-56 Published block diagram showing distribution of human population in the United States in 1960; diagram produced by an IBM 7094 using an offline ink-on-paper CalComp plotter, at the University of Michigan Department of Geography.

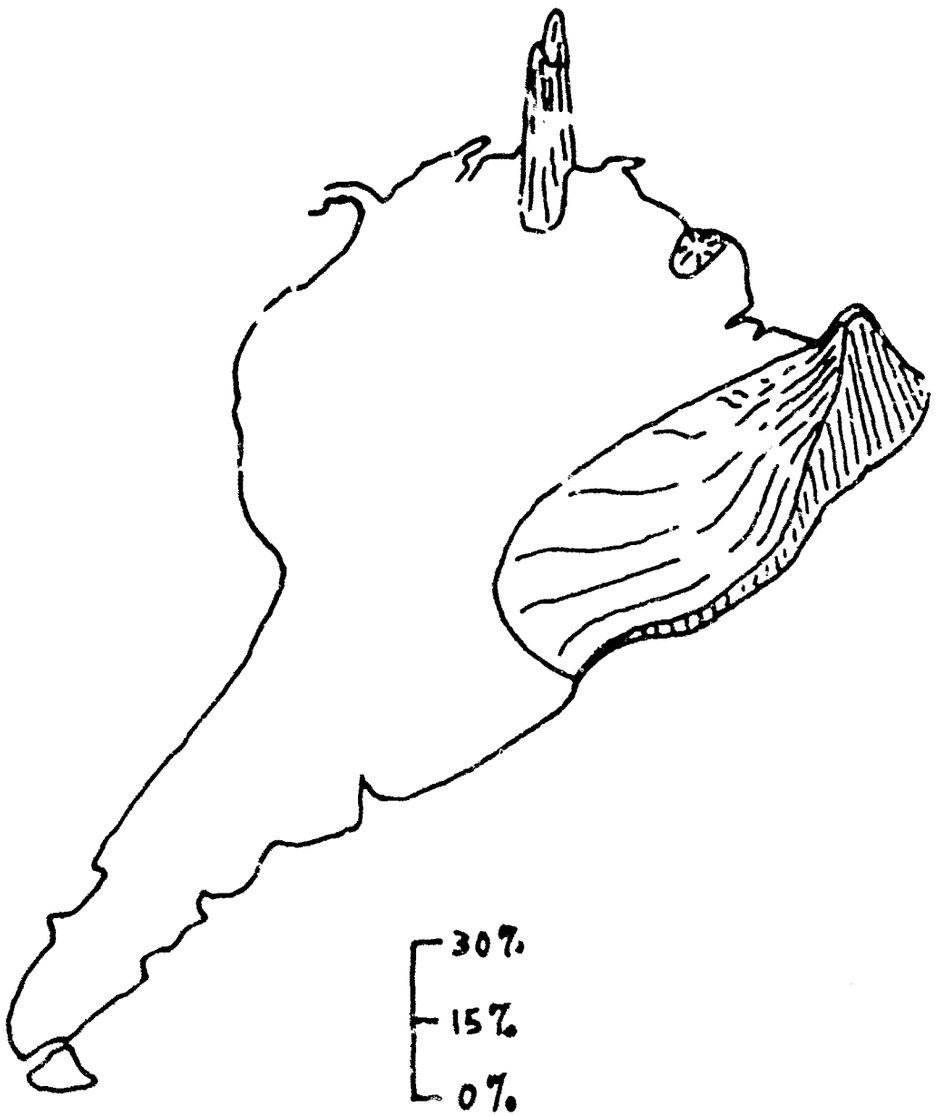


Figure 3-8 The standard test set of MOD schistosomiasis data (Fig. 3-1), presented as a block diagram drawn manually as part of the MOD effort. (The vertical scale inset refers to infection rate.)

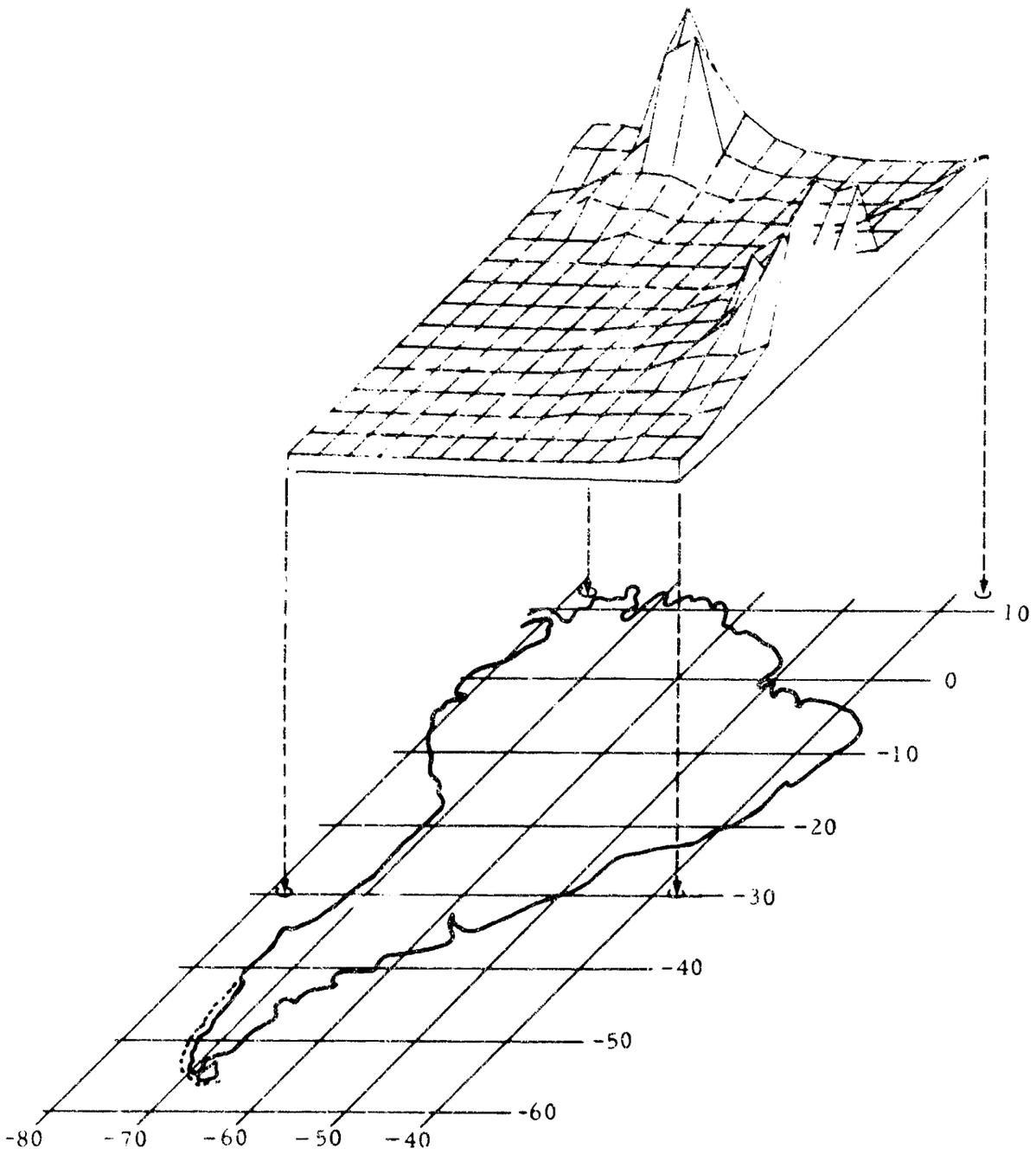
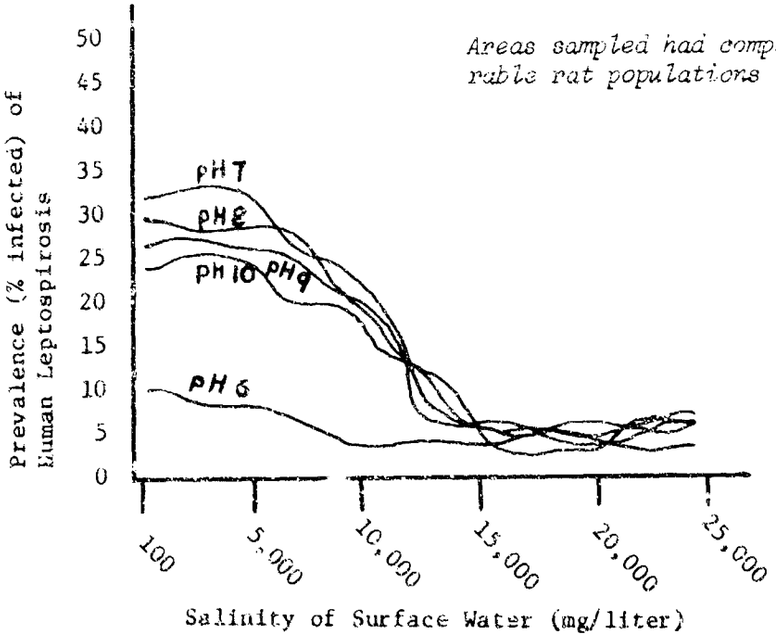


Figure 3-58 A block diagram, presenting the standard set of schistosomiasis test data (Fig. 3-1), produced by an IBM 7094 using an off-line plotter, with the University of Michigan's Department of Geography block-diagram program (continent outline added manually).

We have chosen 100 square miles as a reasonable geographic area to consider in this hypothetical situation, and have (theoretically) reliable data available from most of the area. In 18 of the square miles, because of inadequate human or rat populations, or insignificant surface water, etc., appropriate data could not be collected. Here is the result:



The family of curves shown above represents a <u>hypothetical</u> situation since the data were hypothetical, and this theoretical exercise was simply to illustrate the potential usefulness of a method. If the graph had been based on real data, one could reasonably conclude that:

> *The number of rats in a given area bears a direct relationship to the prevalence of human leptospirosis, but this influence is significant only under the limiting conditions: (1) pH of surface water is within the range of 7 to 10, and (2) salinity of surface water is less than 12,000 mg per liter.*

These data were selected and arranged so that the results would be compatible with our general knowledge of human leptospirosis, however, so far as we are aware, this knowledge has not been supported by data presented and analyzed in the manner that we have described here.

It is known that leptospires do not survive for long (in an infective state) unless surface water is in the general pH range of 7 to 10. There is also evidence to suggest that salinity of surface water influences, independently, survival time of the organisms, but the effects of salinity have not been firmly established. Such a hypothetical study as we have described would go far toward resolving the issue. Other important causal relationships might become evident upon creating additional families of curves in which, for example, the first of the two variables was prevalence of human leptospirosis, and the second, water content of first one specific cation, then others, in turn -- the fixed factor being the pH of water. (In selecting the data for these families of curves, obviously, it would be important to choose geographic areas where rat populations were reasonably comparable.)

3.5 CONCLUSIONS

From our rather detailed analysis of output we have concluded that it is possible to produce meaningful disease-environmental maps by computer, and, furthermore, that such production is feasible. We have pointed out the very considerable limitations of data, and have shown that many data cannot be meaningfully mapped. Furthermore, we have emphasized that the computer system itself cannot perform an analysis of the maps that it produces, but that it can provide useful supporting information, in narrative form, describing certain limitations of the data and providing supplemental and complemental information of a type that cannot be mapped.

Methods of using disease-environmental maps have been considered, and we have described the advantages of keeping these rather simple, and printing them on transparent sheets so that they can be overlaid, one on another

(and on appropriate base maps), and compared with particular concern for coincidence of patterns that would connect various factors involved in the (multifactorial) cause of the disease in question. We believe that visual pattern recognition will continue to be the major method of analyses in most instances because the variation in data coverage and the non-quantitative nature of many of the data will continue (in the foreseeable future) to interfere seriously with rigorous statistical treatment of those data.

The use of block diagrams, to give a "three dimensional" appearance to some aspects of disease-environmental factors, has been described and illustrated, and we have concluded that this is a valuable form of output.

We have considered the application of a particular kind of graph which, when used in a proper series, leads to the production of a "family of curves" that allows a "three dimensional" approach to the problem. This method is, theoretically, a very valuable one but, because of severe demands on the kinds and amounts of data required would probably be most useful in prospective studies that involved relatively small areas.

Possible extensions of the MOD system have been considered, but these considerations have, of necessity, been exploratory. Serious efforts to design some of these methods of extension will not be possible until the system is operational. For example, mathematical models could be designed for purposes of predicting future disease-environmental relationships, given proposed changes in the ecology of a particular area (e.g., the construction of a large dam, or the elimination of a forest, or an extensive program of irrigation converting land to new usage, etc.). It would also be possible to design mathematical models useful in predicting changes in incidence of a given disease based upon past behaviour of the particular disease in a comparable environmental area. Furthermore, matrices of factors could be prepared, analyzing the extent of correlation between each pair of factors. If this approach proved feasible, the computer system itself could compare and analyze data which, otherwise, it would be necessary to output as several different maps. Obviously, such a capability,

even if it were very elementary, could be useful in determining which disease or environmental factors were the most important ones to output in map form.

* * *

Finally -- and most important -- we have concluded that, even though the MOD project has been designed with primary consideration for output of information in map form, as an over-all system, its potential applications go well beyond "mapping of disease". Certain methods which we have developed (particularly in connection with structuring data), and several of the techniques which we have devised, are very pertinent to the structuring and manipulation of many many kinds of data. We believe that, properly used, these methods would be very helpful in converting these data to information.

4 Data characteristics

ABSTRACT - Without an adequate data base no computerized system can function effectively. This section concentrates on the qualities of disease-environmental data in relation to computer manipulation leading to map output. A method of structuring data is developed that provides proper preparation (preprocessing) for computer input. A factor catalogue is developed and types and characteristics of data sources are discussed.

"Bad use of language (in the sense of confusing misuse) usually leads to unresolved controversies."

Anatol Rapoport

4.0 GENERAL CONSIDERATIONS

There is a vast amount of information available pertaining to the ecology of disease, but this information has been directed to such a variety of (discipline-oriented) recipients, that it lacks unifying characteristics.

Not only is there the problem of jargon, i.e., the language peculiar to each discipline, but there is a broader semantic problem in that the same word may have different implications when used by the geographer, the geologist, the agronomist, the limnologist, the parasitologist, the epidemiologist, the veterinarian, the pathologist ..... Furthermore, the structure of (data filled) sentences and phrases differs significantly.

One of the (perhaps the) most important and difficult tasks in developing the MOD system was to design a method by which the pertinent data could be extracted from widely varying source documents and structured (preprocessed) so that they could be input into the computer system in a form suitable for manipulation that would yield mapped (and narrative and tabular) output.

In broad terms we required a data-structuring system that would cut across disciplinary boundries (and foreign language barriers), allowing us to fit various kinds of data into a unifying framework that would provide crisp specifications of:

The thing itself	... WHAT	*QUALITATIVE*
The quality of the thing	... WHAT ABOUT IT	
The value (measure) of that quality	... HOW MUCH	*QUANTITATIVE*
-- and then make this quality/quantity complex mappable by attaching to it a specific location	... WHERE	
	and	
and time characteristic	... WHEN	

A data structuring system that can meet these requirements would suffice for the MOD project, but would have the capability of serving in other areas too. It could handle virtually any sort of data which had a "place" (and time) distribution, e.g., data pertaining to a body or an organ or a tissue, or to a machine, or to a population ("space structured" in terms of age and sex and occupation), or to land usage, or to character and distribution of resources, etc., etc.

Computer technology has progressed to a very sophisticated stage, and the hardware available to the biologically oriented scientist is developed far beyond his capacity to use it The greatest single obstacle to the full-scale effective use of what computer technology has to offer is the lack of such a data structuring system as we have just described. Without effective input there can be no effective storage and retrieval. Storage/retrieval has reached its highest level of (computer) use in relation to various accounting procedures and in maintaining current inventories of material, etc. These are areas in which selection of data to be stored poses no great problem, nor is there any particular difficulty in characterizing the material for ready retrieval since it is already in simple, direct qualitative/quantitative terms.

Since maps are the output upon which the MOD system concentrates, and since maps are ordinarily constructed by selecting a set of data points from which various cartographic representations can be drawn, let us first define maps and data points.

4.1 DATA STRUCTURING TERMINOLOGY

continued next page

(1) MAP:

Definition: A graphic/visual presentation, on a geographic-coordinate basis, of the information imparted by a particular set of specific data points.

(2) DATA POINT:

Definition: A specific geographic locality where a particular factor/aspect/facet of the total disease/environmental situation has been determined/observed/measured, and the result/evaluation/value expressed in some qualitative/quantitative form.

We will return to these two definitions later, after discussing some other necessary ideas.

Note that three fundamentally different concepts are implicit in the definition of a data point: a precise geographic location, a specific value (either a word or a number), and an exact description of the disease/environmental factor involved. In order to understand these three concepts better, let us define and illustrate:

(1) LOCATION (LOC):

Definition: The exact geographic position of the data point, stated as precisely as possible.

Examples: For purposes of the MOD system, the LOC of each data point can be stated in either of two ways:

(a) As the name of a political unit, such as:
Pope County, Illinois, U.S.A., North America, or
Minas Gerais prov., Brazil, South America, or
Bloomington, Monroe County, Indiana, U.S.A., North America.

(b) As a pair of numbers (LO, LA; i.e., longitude, latitude), indicating a point within a particular political unit, such as:
W086°33' N39°07, or
W073°01' N23°23'

Although other methods of stating geographic locations are used, since any geographic location can be expressed in one of the two ways described above, these two are the only ones incorporated into the MOD system.

(2) VALUE (VAL):

Definition: An alphabetic and/or numeric symbol expressing the precise character/condition of that aspect/factor (of the disease/environmental situation) being considered at (the LOC of) the specific data point.

Examples: 0. 1. 2. 3. 0.05. 0-10. 15-35., or Absent. Present. Rare. Common. Abundant. Shale., or Tropical rainforest.

(3) FACTOR:

Definition: Alphabetic and/or numeric symbols naming/describing exactly what part/aspect/facet of the total disease/environmental situation is being evaluated (i.e., given a VAL) at (the LOC of) the specific data point.

Example: Infection rate of schistosomiasis due to Schistosoma mansoni in man during the period 1940 to 1960.

The functions of the three different parts of each data point become quite clear when the process of making a map is considered in detail. First, the cartographer selects from the entire data pool at his disposal a set of data points all of which involve the same aspect of the disease/environmental situation (i.e., which all have the same factor specified). For example, he may select a set of data points all of which concern "total rainfall (inches) during 1966". Second, he plots the position of each data point on a geographic-coordinate grid or base map, utilizing the location (LOC) given for the particular data point. Third, he writes the value (VAL) of each data point next to its plotted position; for example, in mapping "total rainfall (inches) during 1966", he writes the values "43", "56", "17", etc., next to the plotted X's that mark the locations of the data points. Finally, he examines this rough map and draws, around the plotted data points, whichever cartographic symbols he thinks will best present the data.

To summarize, the data point's LOCATION (LOC) describes where a disease/environmental situation was studied. The FACTOR specifies what aspect of the disease/environmental situation was studied. The VALUE (VAL)

gives the result or conclusion reached by the studies. Thus the data point is the combination of a specific location (LOC), a specific factor, and a specific value (VAL).

The concept of "factor" requires more elaboration. Any "factor" is a disease/environmental description of some kind, however, several different types of descriptions can be distinguished. For the purposes of the MOD system we define the following four types of factors:

(1) LOW-ORDER FACTOR (LOF):

Definition: The most specific possible name or description of a particular disease/environmental situation.
Examples: Occurrence. Abundance. Point prevalence. Period prevalence. Incidence. Inches. Leptospirosis. Schistosomiasis. L. pomona. L. canicola. S. mansoni. S. japonicum. Raccoons. Skunks. Foxes. Clinical observations. Isolation from urine. Isolation from tissue. Serologic tests. Mean total annual rainfall. Maximum recorded July rainfall. Savanna. Taiga. Sewer workers. Rice farmers. Cane cutters. Swineherds. Limestone. 1961. 1900-1950.

(2) MIDDLE-ORDER FACTOR (MOF):

Definition: The set of all LOF's which describe the same aspect/facet of disease/environmental situations.

Examples:

MOF	LOF's Making Up The MOF
Measure	Occurrence. Abundance. Point prevalence. Period prevalence. Incidence. Inches.
General kind of disease	Leptospirosis. Schistosomiasis.
Specific disease agent	L. pomona. L. canicola. S. mansoni. S. japonicum.
Animal host infected	Raccoons. Skunks. Foxes.
Method of diagnosis	Clinical observations. Isolation from urine. Isolation from tissue. Serologic tests.

continued next page

Precipitations	Mean total annual rainfall. Maximum recorded July rainfall.
Vegetation	Savanna. Taiga.
Occupational group	Sewer workers. Rice Farmers. Cane cutters. Swineherds.
Bedrock	Limestone. Granite.
Time period for which data applies	1960. 1961. 1962. 1963. 1964. 1900-1950.

(3) HIGH-ORDER FACTOR (HOF):

Definition: A specific combination of LOF's in which each LOF belongs to (is drawn from) a different MOF, i.e., a specific combination of LOF's to which no MOF contributes more than one LOF.

Examples: (In these examples, words standing alone are LOF's; words in brackets "[ ]" are MOF's; and words in italics are connectors.)

HOF #1 = Incidence [measure] *of* leptospirosis [general kind of disease] *due to* L. pomona [specific disease agent] *in* skunks [animal host infected] *as determined by* isolation from urine [method of diagnosis] *during* 1962 [time period for which the data applies].

HOF #2 = Inches [measure] *of* mean total annual rainfall [precipitation] *for* 1963 [time period for which the data applies].

HOF #3 = Abundance [measure] *of* raccoons [animal host infected] *in* taiga [vegetation] *on* limestone [bedrock] *during* 1965 [time period for which the data applies].

(4) POLY-ORDER FACTOR (POF):

Definition: A specific combination of LOF's in which at least two LOF's belong to (are drawn from) the same MOF, i.e., a specific combination of LOF's to which at least one MOF contributes more than one LOF.

Examples:

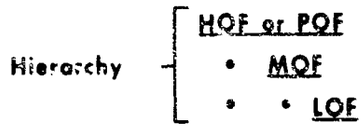
POF #1 = Period prevalence [measure] *of* leptospirosis [general kind of disease] *due to* L. pomona *and* L. canicola [specific disease agent] *in* foxes

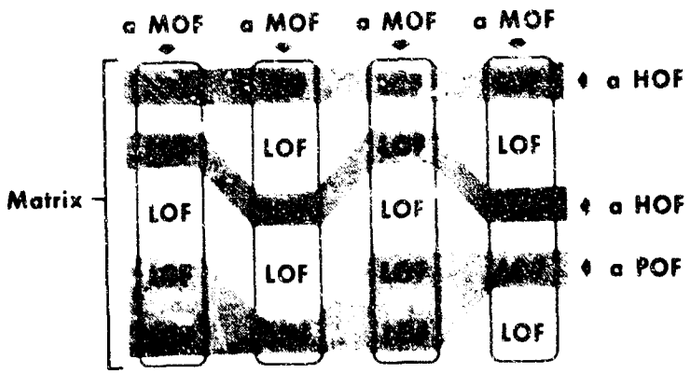
continued next page

[animal host infected] *as determined by* clinical observations, isolation from tissue, and serologic tests [method of diagnosis] *for* 1960, 1961, 1962, 1963, 1964, *and* 1965 [time period for which the data applies].

POF #2 = Inches [measure] *of* mean total annual rainfall [precipitation] *for* 1900-1950 [time period for which the data applies].

LOF's, MOF's, HOF's, and POF's, together, can be viewed either as a kind of hierarchy or as a kind of matrix:





There are two kinds of MOF's:

(1) C-MOF's (common - MOF's), and

(2) O-MOF's (optional - MOF's)

C-MOF's are those MOF's which should, and usually do, accompany (as necessary descriptive elements), or should be common to every data point (i.e., every bit of mappable data), regardless of what aspect of the

disease/environmental situation the data point concerns. However, C-MOF's are not absolutely essential for mapping data; a data point can be plotted/mapped just so long as it has a specific location (LOC), a specific value (VAL), and at least some specific factor (which may or may not include some or all of the C-MOF's). Because C-MOF's should fit into the statement of the factor of every data point, it is desirable to differentiate them from all the other MOF's, i.e., the O-MOF's, which need not fit into every disease/environmental data point (and which in a sense then, are optional). As a rule, O-MOF's will be part of every mappable data point, but both the particular O-MOF's used and their number can vary widely from one mappable data point to another.

Examples of O-MOF's include most of those MOF's given as examples under the definition of MOF's in general; C-MOF's, on the other hand, include these and only these MOF's:

C-MOF	LOF's Making Up The C-MOF
(1) Security classification of the data.	Top secret. Secret. Confidential. Restricted -- for official scientific use only. Unclassified.
(2) Primary source document identification. (The primary source document is the paper which originally reported the data.)	Abbreviated bibliographic citation: author(s), date, journal/book, volume, page. (This MOF will always be used, whether the data comes from primary or from secondary source documents.)
(3) Secondary source document identification. (The secondary source document is a paper which references or quotes data already reported.)	Abbreviated bibliographic citation: author(s), date, journal/book, volume, page. (If the data is being extracted from its primary source document, this MOF will be left blank or not used.)
(4) Professional evaluation of data source (i.e., author, organization, institution, source document, etc.).	More reliable. Less reliable. Reliability not assessed.

• *(See pp. 4 - 34 and - 35 for further basic discussion of data structuring terminology.)*

continued next page

(5) Computer evaluation of data point (to be calculated internally by the computer).	(a number)
(6) Time period for which the data applies.	1963. 1960-1964. June 1965. Pre-1966. 17 April 1964.

LOF's and MOF's by themselves cannot (usually) stand alone as the complete statement of the factor for a set of data points being mapped, because LOF's and MOF's do not, in general, convey sufficient information. On the other hand, HOF's and POF's can (always) be meaningfully mapped, each HOF or POF serving as the complete statement of the factor being mapped or as the description (title or legend) for the map of that factor.

Some idea of the possible number of mappable factors -- HOF's and POF's -- can be obtained from estimates of the total number of LOF's and MOF's available to combine into HOF's and POF's. We estimate that on the order of 10^{30} different HOF's and POF's could be constructed. Obviously, computer handling of this tremendous number of possible factors is not only highly desirable, but a virtual necessity.

In special cases certain elements can be treated either as LOF's or as VAL's. If a specific element is treated as a LOF, that element will appear as part of a particular HOF/POF. If the same element is treated as a VAL, the element will also appear as part of a HOF/POF, but it will be a different one. In other special cases a HOF can contain only one O-MOF, which, in turn, can contain only one LOF. Putting it a different way, sometimes a data-point factor can consist of but a single description or name. Such a factor, in terms of the data structure that we have discussed in detail, can be viewed as a uni-LOF/uni-(O-)MOF/HOF, or, in other terms, as a LOF that is also an (O-)MOF which, in turn, is a (mappable) HOF. In our experience these special cases have been limited to environmental (not disease) situations. As an example of this kind of special case, consider a

data point where the vegetation type is tropical rainforest; the mappable data point could be either:

> In Brazil (LOC), "common" (VAL) *expressing the abundance of* schistosomiasis due to S. mansoni and S. japonicum in man living in tropical rainforests. The factor here is a POF with one O-MOF being "vegetation type", and its contained contributed LOF being "tropical rainforest".)
>
> or
>
> In Brazil (LOC), "tropical rainforest" (VAL) *expressing the vegetation type.* (The factor here is a uni-LOF [vegetation type] uni-O-MOF [biotic communities present] HOF that has the value, "tropical rainforest", at the location being considered.)

Scientific articles and reports often yield narrative-type information that cannot be effectively structured as part of a data point, although the information would contribute to an increased understanding of the data mapped. In terms of data structure we have defined this category as:

NARRATIVE (NAR):

> Definition: Supporting, nonmappable prose/narrative/textual information or data associated with a specific data point.
>
> Example: "Water samples from nearby ponds were lost en route to the laboratory."

To illustrate the use of data structured in this manner we have discussed here are examples of two data points and a map on which they are presented. (In these examples words standing alone are LOF's; words in brackets "[ ]" are MOF's; words in italics are connectors; and words in parenthesis "( )" indicate the major part of a data point.)

> Data Point #1: *At* (LOC:) W076.7° N39.0°, Bowie, Prince Georges County, Maryland, U.S.A., North America, (VAL:) 1/30 *expresses* (factor here, a POF:) period prevalence [measure] *of* leptospirosis [general kind of disease] *due to* L. pomona *and* L. canicola [specific disease agent] *in* foxes [animal host infected] *during* 1960-1965

[time period for which data applies] *according to* Hopps and Kappus, 1967, personal communication [primary source document].

Data Point #2: *At* (LOC:) W077.0° N38.9°, District of Columbia, U.S.A., North America, (VAL:) 1/10 *expresses* (factor, here, a POF:) period prevalence [measure] *of* leptospirosis [general kind of disease] *due to* L. pomona *and* L. canicola [specific disease agent] *in* foxes [animal host infected] *during* 1960-1965 [time period for which data applies] *according to* Richmond and Coffey, 1966, J. Leptospirology, v. 39, p. 107 [primary source document]; (NAR:) the majority of foxes examined in this urban area were in the National Zoological Park.

A map presenting these two data points is shown in Fig. 4.1.

The guiding principle throughout this description of data-structuring methods/techniques of the MOD system has been to relate all concepts to their ultimate application -- the construction of maps synthesized from specific sets of data points. Consequently, attention has been focused on the various component parts of a data point.

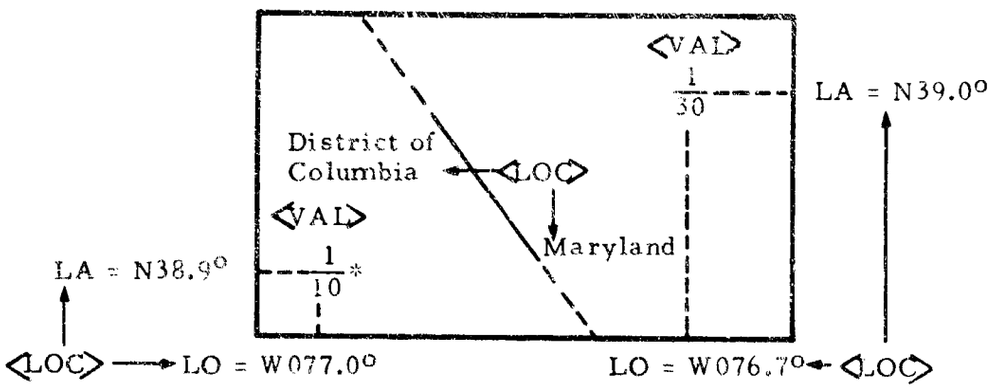
DATA POINT

(1) a location (LOC) *
(2) a value (VAL)
(3) a factor (HOF or POF, *made up of LOF's belonging to MOF's)*
(4) a narrative (NAR) -- *this is not mandatory*

---

* Parenthetically, as will be discussed later, the MOD data processing system treats LOC, VAL, NAR, and all MOF's in essentially the same manner.

<Factor, here a POF:> Period prevalence [measure] (of) leptospirosis [general kind of disease] (due to) L. pomona and L. canicola [specific disease agent] (in) foxes [animal host infected] (during) 1960-65 [time period for which data apply] (according to) all sources [primary source document]:



<NAR> * Foxes here tend to be scared away by numerous tourists, so that the value determined may be low.

Figure 4-1 Map constructed from illustrative data points #1 and #2, given on preceding page, to show the function of various portions of the structured MOD data.

4.2 FACTOR CATALOGUE

4.2.1 DISEASE FACTORS

Disease data sought for input to the MOD system can be structured according to the entries in this section. As previously described, the data must be cast in the form of distinct data points, each containing a geographic location (LOC), value (VAL), factor statement, and supporting narrative (NAR). The factor statement is constructed by combining LOF's from many of the MOF's specified herein.
Throughout this section the following format is utilized:

(MOF code designation) -- MOF name or textual description.

Examples of LOF's (names or textual descriptions) belonging to this particular MOF

Explanatory comments regarding this particular MOF

The precise format used here meets MOD data-collecting and data-processing requirements, as discussed extensively later in this report. The MOF's in this section are listed alphabetically by their 3-letter code designations so that this catalogue may serve as a guide for MOD data extractors.

Almost every MOF described herein can have as LOF's, "other" and "unspecified/undifferentiated" -- consequently, these particular LOF's are not listed under each MOF. Furthermore, every MOF can have "unknown" as a LOF. Because of the manner in which MOD data processing will be performed, this particular LOF is entered implicitly simply by omitting the MOF in question from the data-point factor statement, i.e,, not filling in the appropriate blanks on MOD data extraction forms.

The MOD disease factors are as follows:

(AID) Domestic status of animal infected.

Domesticated.
Wild (feral).
Wild (native).

(AIP) Precise identity of animal infected.

Homo sapiens.
Man.
Dog.
Felis leoparda.
Microtus pennsylvanicus (meadow vole).
Raccoons.
Stink-pot turtle.
Bears (ursids, Ursidae).
Mice (suborder Myomorpha).
Culicinae (culicine mosquitoes).

Allowance must be made for over 2,000 LOF's, which will be arranged in a hierarchical tree-structure.

(AVC) Average course of disease in this outbreak.

Acute (i.e., sudden onset, sharp rise, short duration).
Subacute.
Subchronic.
Chronic (i.e., gradual onset, gradually progressive, long or indefinite duration, or frequently recurring).

(AVD) Average duration of disease in this outbreak.

7 days.
153 days.

(AVS) Average severity of disease in this outbreak.

Fatal.
Severe clinical.
Moderate clinical.
Mild clinical.
Asymptomatic / subclinical.

(CEN) Computer evaluation of data point (to be calculated automatically by the MOD computer system; a C-MOF).

3.
15.
21.

(CRD) Domestic state of carrier.

Same as (AID).

(CRP) Precise identity of carrier.

Same as (AIP).

(No code designation) Data point number.

661223RJC039.
670815JHC173.

This number is not actually a "factor" in the sense in which the MOD data structure was defined previously, but it is necessary for MOD data-processing operations. No code designation is required because these characters, by themselves, will signal the beginning of a new data-point record within the MOD system. Each data point must have a unique designation; this requirement is satisfied by entering a different number on each data extraction form completed. The number consists of the last two digits of the year, two digits for the month, and two digits for the day on which the data form was filled out, followed by three initials of the data extractor's name, then by three digits which indicate that the data point is the nth point extracted by that particular extractor on that specific date.

(DMS) Disease measure (method of indicating extent of disease within population, i.e., "epidemiological index").

Occurrence (necessitates VAL's of "present/absent").
Abundance (necessitates VAL's of "absent/rare/common/abundant").
Point prevalence.(Dorn, 1957).
Period prevalence. (Dorn, 1957).
Incidence. (Dorn, 1957).
Mortality.
Standardized mortality ratio. (Howe, 1963)
Infection rate.
Case rate.
Attack rate.
Hospital-admission rate.

(necessitate VAL's which are ratio, fraction, or percentage numbers, e.g., 1:4, 1/250, 0.07, or 13%)

continued next page

Number of cases existing at specific point in time (i.e., a day) (=CP).
Number of cases existing at any time during specific time interval (i.e., a week, month, or year), including both those which began before and those which began after start of time interval (=CBA)
Number of cases beginning during specific time interval, including only those which began after start of time interval (= CA).
Number of deaths occurring during specific time interval (= DI).

(necessitate VAL's which are absolute numbers, e.g., 15, 749, or 136).

The disease measures requiring ratio/fraction/percentage numbers as VAL's, such as point prevalence, period prevalence, incidence, etc., are so frequently utilized imprecisely and ambiguously in the literature that it may be necessary (with some groups of data) to synonymize them simply as "Morbidity". Similarly, the various measures which are stated as numbers of cases may also have to be synonymized as simply "Number of cases". At this time, however, the definitions given by Dorn (1957) will be followed for point prevalence (=CP/P), period prevalence (=CBA/P), incidence (=CA/P), and mortaligy (=DI/P), where P = population examined. Standardized mortality ratios are calculated from standard health statistics by several arithmetic manipulations as outlined by Howe (1963, pp 3-6).

(DOR) Duration of outbreak reported.

6 days.
30 days

(ESD) Epidemiologic state of disease within population.

Endemic/enzootic (i.e., disease continuously present at low rate).
Hyperendemic/hyperenzootic (i.e., disease continuously present at high rate).
Sporadic (i.e., disease intermittently present, only at low rate).
Epidemic/epizootic (i.e., disease intermittently present, at high rate, in a small area -- or disease continuously present but now at much higher rate than usual, in a small area).
Pandemic/panzootic (i.e., disease intermittently present, and at high rate, over very wide region -- or disease continuously present but now at much higher rate than usual, over very wide region).

(FOP) Frequency of outbreaks preceding outbreak reported.

No outbreaks previously reported.
Outbreaks rare/occasional/seldom.
Outbreaks common/frequent.
Outbreaks very common/very frequent.

(GKD) General kind of disease.

Leptospirosis (= 7-day fever = Weil's disease/syndrome = Ft. Bragg fever = spirochetal jaundice = pea-harvest fever = water fever = (sugar-)cane fever = rice-field fever = harvest fever = swamp fever = swineherd's disease = mud-harvest-field fever = brushy creek fever = pretibial fever = field fever = mud fever = autumnal fever).
Schistosomiasis.
Rabies.
Malaria.
Hemorrhagic fever.
Dengue.
Cholera.

(HSC) Type of human settlement where outbreak occurred.

Urban or large city.
Suburban area.
Small town.
Densely-settled rural area.
Sparsely-settled rural area.

(IHD) Domestication of intermediate host.

Same as (AID).

(IHP) Precise identity of intermediate host.

Same as (AIP).

(IMU) Prior state of immunity of animals infected in this outbreak.

Susceptible/not immune.
Naturally immune.
Artificially immunized.

(KOR) Kind of outbreak reported.

Isolated case (1).
Small group of cases (2-29).
Large group of cases (30+).

(LBS) Basis of sampling for largest sample involved (i.e., how were the individuals chosen for examination or testing, in the broadest sense).

Random survey of townspeople.
Sick people visiting local physicians for any illness.
Patients admitted to hospital for suspected leptospirosis.
Complete survey of military inductees.
Complete survey of dairy herds.
People living on the same side of a street.
Persons of a particular occupation.

(LDO) Lethality of disease in this outbreak.

Always fatal.
Often fatal.
Seldom fatal.
Rarely fatal.
Never fatal.

(LGD) Relative level of generalization of data.

Data very generalized and broad.
Data intermediate in generality.
Data highly specific.

(LOC) Geographic location of data point (i.e., where did the cases occur?).

North America, United States, Pennsylvania, Centre County, Pine Grove Mills.
North America, United States, Maryland, Anne Arundel County, W 076° 29' N 38° 59'.
Europe, France, Dordogne, _____, Les Eyzies.
Africa, Ghana, Accra Prov., _____, A.E.Prince's lot near Teshi.

This is not actually a "factor" in the sense in which the MOD data structure was defined previously, but it is necessary for MOD processing operations. Allowance must be made for over 10,000 different entries, entries which will be arranged in a hierarchical tree-structure.

(LOG) Occupational groups in largest sample involved.

Sewer workers
Rice-field farmers.
Cane-cutters.
Butchers.
College/university students.

(LRE) Racial/ethnic/breed groups in largest sample involved.

Scandinavian Europeans.
Nilotic Negroes.
Italian Americans.
Hereford cattle.
Chihuahua dogs.
Tagalog tribespeople.

(LSA) Ages in largest sample involved.

Adolescent.
Adult.
22-28 yrs.
0-6 mos.

(LSX) Sexes in largest sample involved.

Male
Female.

(LSZ) Size of largest sample involved (largest number of individuals examined, i.e., how many individuals were examined or tested in the broadest sense?).

703.
514.
37.

(MDG) Method of diagnosis.

Clinical observation.
Isolation of disease agent, source unspecified.
Isolation of disease agent from water.
Isolation of disease agent from soil.
Isolation of disease agent from urine.
Isolation of disease agent from blood.
Isolation of disease agent from tissue.
Serologic test -- single specimen (or non-rising titre).
Serologic test -- multiple specimens -- rising titre.
Skin test.
Xerodiagnosis.
Biopsy.
Autopsy.

(MFD) Medical facilities involved in diagnosis.

Military hospital/clinic.
University/academic hospital/clinic.
Large/urban hospital/clinic.
Small/rural hospital/clinic.
Individual physician.
Nurse/paramedical person.
Folk/witch doctor.
None.
Roving expedition / field study.

(MFT) Medical facilities involved in treatment during this outbreak.

Same as (MFD).

(MND) Minimum duration of cases in this outbreak.

Same as (AVD).

(MNS) Minimum severity of disease in this outbreak.

Same as (AVS).

(MRG) Manner of reporting/grouping data.

Data not grouped -- reported as individual cases.
Data grouped and reported by county.
Data grouped and reported by state/province.
Data grouped and reported by country/colony/dependency.

(MTR) Method of transmission of disease to animal infected.

Direct contact with living infected animals.
Direct contact with dead animal,tissue, or blood.
Direct contact with excreta (including urine).
Indirect occupational contact with water.
Indirect recreational contact with water.
Indirect domestic contact with water.
Indirect occupational contact with soil.
Indirect recreational contact with soil.
Indirect domestic contact with soil.
Bite of carrier or vector.

(MXD) Maximum duration of cases in this outbreak.

Same as (AVD).

(MXS) Maximum severity of disease in this outbreak.

Same as (AVS).

(NAR) Supporting narrative information for data point.

Water samples taken during expedition were contaminated accidentally before laboratory processing.
Observations within this tract of jungle led to a mapping survey several months later.

This is not actually a "factor" in the sense in which the MOD data structure was defined previously, but it represents significant information that can be made available to the MOD user.

(PES) Professional evaluation of data source (i.e., author, organization, institution, source document, etc.; a C-MOF).

More reliable.
Less reliable.
Reliability not assessed.

(PFD) Pattern of fluctuations in disease situation over long periods of time.

Disease continuous, with no significant peaks (peak = 3 x average value).
Disease continuous, but with significant peaks.
Disease intermittent and seasonal, within a year.
Disease intermittent and non-seasonal, but within a year.
Disease intermittent and seasonal, but non-yearly.
Disease intermittent and non-seasonal, and non-yearly.

(PHD) Domestic state of primary host.

Same as (AID).

(PHP) Precise identify of primary host.

Same as (AIP).

(PSD) Primary source document identification (abbreviated bibliographic citation; a C-MOF)

Smith & Jones, 1968, J. Zool, v. 13, p. 51.
Kappus, 1968, Leptospirosis in Mice, p. 157.

continued next page

Hopps, 1966, Princ. of Path., pp. 13-17.
Brown, 1951, Geol. Bull., v. 51, pp. 571-703.
Richmond. 15 Feb 67, MOD System & Data Analysis Results (Final Report from PRC), p. 13.

The primary source document is that in which the data was originally reported.

(RSD) Domestic status of reservoir.

Same as (AID).

(RSP) Precise identity of reservoir.

Same as (AIP).

(SBS) Basis for sampling for smallest sample involved (i.e., how were the individuals chosen for examination or testing in the narrowest sense?).

Same as (LBS).

(SDA) Specific disease agent.

Leptospira, serotype not determined or not specified.
Leptospira, all serotypes.
Leptospira pomona.
Leptospira canicola.
Schistosoma mansoni.
Plasmodium falciparum.
Vibrio cholerae.
Omsk hemorrhagic fever (OHF) -agent.

Agents are to be arranged in a hierarchical tree-structure, with allowance for about 125 leptospiral agents and about 80 hemorrhagic-fever agents.

(SEA) Season for which data applies.

Winter.
Spring.
Summer.
Fall.
Deer-hunting season.
Wet season.
Dry season.

(TOD) Time of day for which data applies.

Morning-night.
Dawn.
Morning.
Mid-day.
Afternoon.
Dusk.
Evening-night.
Mid-night.

(TSC) Topographic situation of cases in outbreak.

Along river/valley bottoms.
On plateaus.
On mountains, just below tops.

(UDS) Unique designator for combination of smallest and largest samples involved.

670502RJC073.
680923HCH107.

The matter of sampling to determine the value of a data point is extremely complex, and is best explained by using a simple example.

Assume that, in a town of 10,000 people, 1,000 men have been inducted into the military -- 800 of these have been examined medically in some way or other, 500 of which have been examined serologically for leptospirosis; 10 are found to have L. pomona antibodies, and 5 are found to have L. canicola antibodies, and 2 are determined to have leptospirosis, clinically, but were not tested serologically.

There are actually three separate data points which must be extracted separately from the above situation:

(1) 10/500 point prevalence of L. pomona serologically.
(2) 5/500 point prevalence of L. canicola serologically.
(3) 2/800 or, possibly, 2/1000 point prevalence of L., serotypes undertermined, diagnosed clinically.

Commonly, when such studies are published, one or more of the figures: 1000, 800, and 500, are omitted. Even though they were given, the fact that all these numbers relate to essentially the same sampling situation might become obscured if the data points were entered into the MOD system separately. To avoid this, a unique

designator is required for the combination of samples (the 1000, 800, and 500) involved in this situation. Such a unique designator can be constructed from the date (two digits for year, two digits for month, and two digits for day), three letters for extractor's initials, and three digits indicating that this was the nth sample encountered by the extractor on that day.

The unique designator is required for computerized combination of these data points. For example, given the same unique sample designator for the 3 data points above, the computer system can then combine them correctly to get 10 + 5/500 point prevalence of L. serologically (rather than 10 + 5/500 + 500), and 10 + 5 + 2/800 to 10 + 5 + 2/1000 point prevalence of L. serologically and/or clinically (rather than 10 + 5 + 2/500 + 800 to 10 + 5 + 2/500 + 500 + 1000), in this particular leptospirosis situation.

(No code designation) Security classification of data (a C-MOF).

Top secret.
Secret.
Confidential.
Restricted -- for official scientific use only.
Unclassified.

No code designation is required, because security matters must be handled manually rather than by the MOD computer system itself.

(SOG) Occupational groups in smallest sample involved.

Same as (LOG).

(SRE) Racial/ethnic groups in smallest sample involved.

Same as (LRE).

(SSA) Ages in smallest sample involved.

Same as (LSA).

(SSD) Secondary source document identification (abbreviated bibliographic citation; a C-MOF).

Same as (PSD).

The secondary source document is one which references or quotes data already reported elsewhere.

(SSX) Sexes in smallest sample involved.

Same as (LSX).

(SSZ) Size of smallest sample involved (smallest number of individuals examined, i.e., how many individuals were examined or tested in the narrowest sense?).

Same as (LSZ).

(TGA) Treatments given to animals infected.

300,000 units K phenoxmethyl penicillin q.i.d. for 7 days.
Nodules surgically removed.

(TIM) Time period for which the data applies (i.e., when did the cases occur?; a C-MOF).

63 (i.e.,1963).
6306,6411 (i.e.,Jun 1963 - Nov 1964).
630617 (i.e.,17 Nov 1963).

(VAL) Value for data point.

1.
33.
Absent.
Rare.
Tropical rainforest.
Probably present.

This is not actually a "factor" in the sense in which the MOD data structure was defined previously, but it is necessary for MOD processing operations. The particular value entered for a data point must be checked to insure that it is compatible with the factor statement (particularly the disease measure -- DMS).

(VCD) Domestic status of vector.

Same as (AID).

(VCP) Precise identity of vector.

Same as (AIP).

4.2.2 ENVIRONMENTAL FACTORS

Environmental data sought for the MOD system can also be cast in terms of the MOD dat structure, utilizing data points consisting of LOC (location), VAL (value , a factor statement (constructed by combining LOF's from many of the MOF's listed), and N R (narrative), and these factors for environmental data are the same as previously described for diseas data.

Other (0-) MOF's can be constructed according to the same format as diseas -related M 's. For example:

(DMZ) Biogeographic distribution measure.

Occurr ce (necessitates VAL's "absent/present").
Abunda ce (necessi ates VAL's "absent/rare/common/abundant").
Number of individuals seen (necessitates VAL's such as "137").

(SMP) Precise identity of small mammal considered.

Same as disease MOF (AIP).

Because of the enormous number of possible environmental MOF's, this catalogue of environmental factors lists only broad statements as to the kind of data desired rather than precise MOF's. However, precise MOF's can be readily formulated from this catalogue when requirements for specific environmental data arise. Physical/chemical-environmental factors are listed first, then biologic-environmental factors and, finally, human-environmental factors.

The MOD environmental factors are as follows:

Bedrock:
- Types (granite, limestone, schist, etc.)
- Structure (flat-lying, folded, block-faulted, etc.)
- Chemical/mineral content (including major elements, trace elements, etc.)

Soil:
- Types (latosol, sierozem, podzol, etc.)
- Chemical/mineral content (including major elements, trace elements, pH, etc.)
- Temperature (at surface, at 1-foot depth, etc.)
- Moisture content, porosity, permeability

Erosion:
- Severity
- Type (sheet, gully, etc.)

Topography:
- Elevation/altitude
- Relief
- Slopes (abundance, steepness, orientation)
- Landforms (mountain, valley, plain, etc.)

Water:
- Availability (especially of potable water)
- Types (soil water, surface water, ground water, etc.)
- Physical/chemical characteristics (including pH, salinity, temperature, turbidity, hardness, major and trace elements -- especially oxygen content and carbonate content)

Surface water bodies:
- Type (intermittent stream, permanent lake, permanent spring, etc.)
- Origin (natural, artificial -- reservoir, irrigation ditch, broken bottle, etc.)
- Water movements (flowing, stagnant turbulent, etc.; wave direction and height; current direction and speed; tides)
- Hydrography (depth, gradient, bottom type, etc.)
- Biotic content (aquatic weeds, oyster beds, etc.)

Water pollution:
- Type (industrial-chemical, suspended solids, thermal, sewage, etc.)
- Intensity/severity
- Duration/frequency (continual, seasonal, occasional, etc.)

Evaporation, evapotranspiration, dessication (potential, actual)

Climate types (humid mesothermal, humid continental, etc.)

Weather types

Air temperature:
- Where measured (1 inch above ground, 30 inches above ground, etc.)
- When measured (noon, 6 P.M., any random time, etc.)
- Time period involved (daily, monthly, seasonally, annually, etc.)
- How measured (mean, highest observed, lowest observed, range of variation, etc.)

Precipitation:
- How measured (total, range of variation, mean, maximum, etc.)
- Time period involved (daily, weekly, monthly, annually, etc.)
- Seasonal distribution (continually wet, distinct wet/dry seasons, etc.)
- Types (rain, snow, sleet, hail, etc.)

Dew:
- Frequency of formation
- Duration into daylight hours after formation

Frost:
- Frequency of formation (number of frost-free days, etc.)
- Seasonal distribution (date of last spring killing frost, date of first fall killing frost)

Glaciers

Humidity (relative, absolute, wet-dry bulb temperatures, etc.)

Barometric pressure

Clouds and fog, and clarity/transparency of atmosphere:
- Type (dense fog, stratocumulus, etc.)
- Frequency of occurrence

Illumination/light/insolation:
- Days of sunshine
- Length of daylight (also length of growing season)
- Extent to which people are exposed to sun each day

Winds:
- Direction
- Speed/severity/force
- Frequency
- Seasonal distribution
- Special types (up-valley wind, ocean breeze, etc.)

Thunderstorms and lightning (static electricity)

Natural disasters (hurricane, tornado, flood, dust-/sand-storm, drought etc.)

Air polution:
- Type (pollen, smoke, toxic gases, etc.)
- Frequency
- Severity/intensity

Gravity

Magnetism (terrestrial)

Background radiation (ionizing):
- Terrestrial (from uranium-bearing black-shale bedrock, etc.)
- Solar
- Cosmic - ray

Organisms occurring in same area as disease cases (including wild and domesticated; including vertebrate animals, invertebrate animals, plants, and protists; including potential and known intermediate hosts, accidental hosts, artificial or experimental hosts, reservoirs, carriers, vectors, parasites, etc.)

Biogeographic distributions of such organisms:
- Occurrence
- Relative abundance
- Population size and density
- Natural population cycles, variations, & migrations
- Degree of concentration versus dispersal

Living habits of such organisms:
- Feeding (including biting preferences, food chains, etc.)
- Breeding
- Resting (hibernating, aestivating, etc.)
- Competitive and symbiotic relationships
- Amount of contact with man

Disease/health conditions of such organisms

Pesticide or drug resistance among such organisms

Local habitats (grassland, swamp, desert, forest, cultivated field, pasture, etc.)

Biotic communities (short-grass prairie, oak-hickory forest, etc.)

Biomes (tropical rainforest, taiga, etc.)

Biogeographic regions (Neotropical, Holarctic, etc.)

Human population:
- Total numbers of people
- Density (number of people/square mile)
- Rate of increase or decrease of total population
- Birth rate (including effects of birth control measures)
- Death rate
- Population movements (war refugees, nomads, migratory workers, patients admitted to hospitals far from their residences, military troop movements, etc.)

Age structure of population

Sex distribution within population

Racial groups within population

Ethnic or nationality groups within population

Language groups within population

Socio-economic groups within population (including caste)

Blood-group distribution

Distribution of other human hereditary or genetic factors

Personal medical, hygenic, and sanitary practices and habits (washing hands, taking sauna baths, protecting newborn infants, etc.)

Public health service practices, expenditures, and facilities:
- Source of potable water (surface reservoir, drilled well, etc.)
- Treatment of water supply (chlorination, filtering, etc.)
- Treatment of sewage
- General level of community sanitation
- Pest control and eradication programs
- General vaccination or inoculation programs

Medical facilities available:
- Size and type (large hospital, small clinic, mobile aid station, etc.)
- Sponsorship and administration (government, military, missionary, industrial, academic, research institute, private, etc.)
- Numbers
- Availability of pathologist and/or laboratory diagnostic service
- Ease of access to facilities among different groups within population (due to cost, distance, etc.)

Medical personnel available:
Type (physicians, veterinarians, nurses, technicians, dressers, etc.)
Numbers

General health level of population

Other diseases common in population (including genetic defects, infectious diseases, mental disorders, malnutrition, alcoholism, drug addiction, etc.)

Nutritional and dietetic habits and customs:
Physical/chemical/mineral content of foodstuffs
Nutritional deficiencies

Settlement patterns (urban, suburban, small town, dense rural, sparse rural, etc.)

Housing preferences and habits:
Length of residence in presently-occupied dwelling
Construction (windows screened, walls brick, roof straw, etc.)
Number of people living in each house
Amount of time spent indoors

Types and sizes of family groupings

Marriage and divorce customs

Personal clothing habits

Recreational, entertainment, and social habits:
Kinds (swimming at public beaches, etc.)
Frequency
Types involving special risk of exposure to disease (water sports, hiking in jungle, eating raw fish, etc.)

Educational level:
Literacy
Number of high-school, college, etc. graduates
Educational facilities (schools) available

Land use (grazing, farming, reclamation projects, etc.)

Type of economy (hunting-gathering, farming, machine civilization, etc.)

Basis of economy:
Hunting or gathering
Fishing

continued next page

Forestry
Agriculture (farming, ranching, etc.; crops involved, e.g., cotton, wheat, sesame, etc.; agricultural practices, such as use of irrigation, chemical fertilizers, human nightsoil, etc.)
Mining (coal, iron, copper, sulfur, diamonds, petroleum, etc.)
Manufacturing (principal raw materials, principal products)
Services (teaching, research, consulting, etc.)

Occupations and jobs:
Types present (automobile mechanic, university professor, etc.)
Relative proportions (i.e., predominant jobs)
Kinds involving special risk of exposure to disease in which interested (such as butchers, sewer workers, etc.)

Economic levels, distribution of income, standard-of-living index, unemployment

Communications available:
Types
Extent to which utilized

Transportation available:
Types (railroad, private car, bus, etc.)
Extent to which utilized
Mobility and travel pattern of population
Kinds involving special risk of exposure to disease in which interested (such as walking through jungle, fording streams, etc.)

Crime statistics

Military organization of population (none, militia, away-from-home active duty, etc.)

Political movements, political views

Religions and religious/superstitious customs (such as pilgrimages, washing in rivers with other worshippers, etc.)

Artistic, musical, and literary customs and activities

The MOD method of data structuring is, admittedly, hard to grasp. In an attempt to simplify explanation of the basic concept -- and the method of its application -- we conceived of the analogy shown in the adjacent figure.

Figure 4-2 presents an orchard that consists of a number of trees; similarly, a map consists of (is drawn from) a number of data points, and, in our illustration, a single tree is analogous to a data point.

The location of the tree within the orchard is comparable to the location (LOC) of the data point. The size of the tree may be considered comparable to the value (VAL) of the data point.

Carrying our analogy further, the various parts of the tree can be compared with the various parts of the factor statement. The most obvious, vital, specific items in an orchard (from the grower's viewpoint), are the fruit on each tree; analogously, the most obvious, vital, specific items in the data point's factor are the individual LOF's. Furthermore, the branches of the tree (bearing fruit/LOF's) can be compared with MOF's, and the trunk (supporting the fruit-bearing branches/LOF-bearing MOF's) can be compared to a HOF/POF.

As a second analogy, consider the situation in ordinary computerized information-processing terminology. Usually, terms are defined in two levels -- as descriptors and as elements. Elements are analogous to MOF's, whereas descriptors of those elements are equivalent to LOF's.

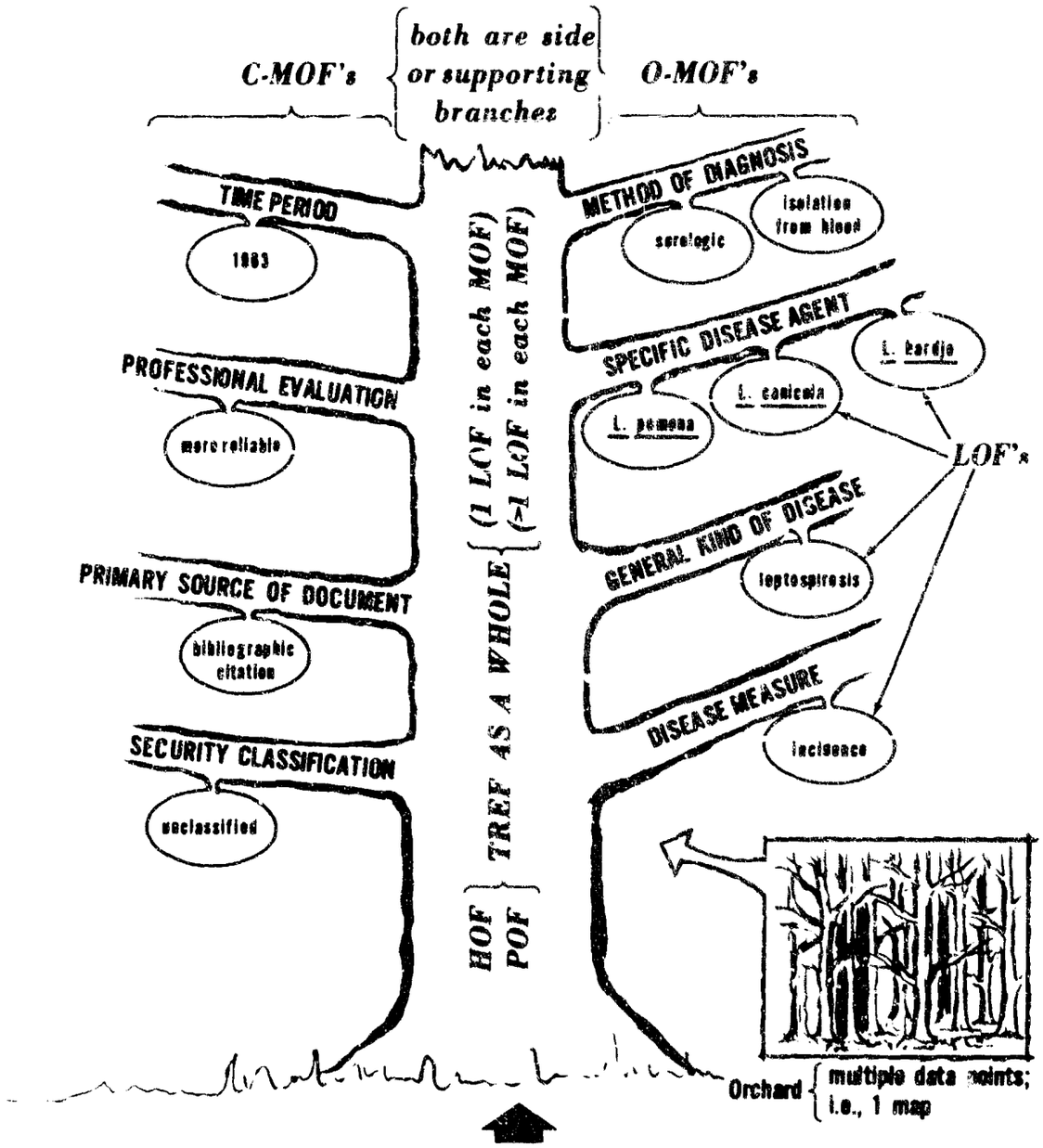


Figure 4-2 Illustrating the analogy between MOD data structure and an orchard.

4.3 DATA REQUIRED FOR MAPPING -- MINIMAL AND OPTIMAL

Now that the structure of the data available to the MOD system has been described, we can more clearly delineate precisely what data are necessary for mapping disease-environmental factors -- and what form they must be in. First and foremost, the data must be capable of being put into the format of discrete data points; and this means that they *must* have a defined (stated) geographic location (LOC), a value (VAL), and some statement of factor (see pages 4-6 - 4-11).

If data were collected to meet a particular need, e.g., a specific MOD system use, its format could be rigidly defined beforehand, and many problems would be avoided. Unfortunately, the MOD system user will, as a rule, be dependent upon data which he had no role in generating -- data that were developed without any thought of computer processing, much less mapping.

Although "geographic location" and "value" are not without their problems, the greatest difficulties come with FACTOR. In part this reflects the vagueness of our language, in part the enormous number of attributes which characterize medical environmental situations, many of which are incomplete in themselves, many of which overlap. It is factor, more than any other component of the data point, that restricts usefulness of the data. It is a limited specification of factor in the original data source that is most likely to "handicap" the data point so that it is closer to minimal than optimal. This is why the consideration of minimal and optimal data required for mapping concentrates on factor.

In data points which deal with certain environmental situations it is possible for the factor to be adequately stated by a single LOF (as described earlier in discussing the terminology of the MOD data structure).

Such a factor is called a uni-LOF uni-(O-)MOF HOF. But with data points that deal with a disease situation, the factor statement requires a combination of several LOF's/MOF's. Our experience indicates that LOF's from the following six MOF's must be given in order for any disease data point to be mappable.

(1) Time period for which data applies *(whether this is the date of onset, or date of termination of disease, or some intermediate time during its course).*

(2) Disease measure *e.g., "number of cases existing during the specific time interval".*

(3) General kind of disease, *e.g., "diarrhea" or "leptospirosis".*

(4) Disease agent (as precise as possible), *e.g., "Schistosoma mansoni" or "Leptospira pomona".*

(5) Method of diagnosis, *e.g., "serologic" or "isolation from urine".*

(6) Identity of animal infected (as precise as possible), *e.g., "Homo sapiens" or "Canis domesticus".*

These requirements are not so absolutely restrictive as they may seem, however, the less specific the factor statement is with respect to any one of these points, the less useful the data becomes. To consider some very bad examples:

> Time period might have to be specified as "some time during 1900-1950"; General kind of disease as "diarrhea"; Disease agent as "unknown"; Method of diagnosis as clinical impression; Identity of animal infected as "nonhuman animals".

These six items represent the minimal requirements for a meaningful disease-environmental data point. Further characterization of the data point's factor would permit the data point to be used in some of the more complicated MOD system calculations. Consequently, it would be highly desirable to include, in the factor statement, LOF's belonging to the following MOF's:

(1) Primary source document. *

(2) Secondary source document. *

(3) Professional evaluation of data source. *

(4) Computer evaluation of data point. *

(5) Unique designator for combination of smallest and largest samples involved.

(6) Size of smallest and (7) largest sample involved.

(8) Age group of smallest and (9) largest sample involved.

(10) Sex of smallest and (11) largest sample involved.

(12) Racial or ethnic group of smallest and (13) largest sample involved.

(14) Occupational group of smallest and (15) largest sample involved.

(16) Basis for sampling of smallest and (17) largest sample involved.

(18) Domestic state of (nonhuman) animal infected.

(19) Epidemiologic state of disease within population of animal (including human) infected.

Many other disease-related LOF's and MOF's can be devised, as demonstrated in the catalogue of MOD system factors and on the data extraction forms (to be described later). Any or all of these can be added to a data point as part of the statement of the data point's factor. However, the six "required" MOF's that we have listed, plus the LOC and VAL, comprise the elements minimally required for effective processing of the disease data point by the MOD computerized mapping system. An optimal data point would include the nineteen "highly desirable" MOF's listed -- and many more, selected on the basis of known present need together with estimates of probable future requirements. But until there has been considerable operational experience with the MOD system, it will not be profitable to

---

* These are C-MOF's and, as previously stated, should accompany every data point, however, they are not required in order for the data point to be mappable.

speculate what precise combination of MOF's/LOF's make up the factor of the "optimal" data point.

On the basis of rather limited experience we have found that disease maps were more readily constructed, and imparted more information to the viewers, if the data points had the following characteristics:

(1) Their values (VAL's) were all originally given in consistent, uniform format, so that no interpretative equating of dissimilar types of values had to be performed.

(2) Their factors were comparatively short, simple, and straightforward, rather than complex and involved. (This trait enables a number of points to be compared with reasonable assurance that each point describes the same aspect of the disease/environmental situation.)

(3) Their factors were oriented toward the raw data rather than toward the conclusions drawn by authors of the papers from which the points were extracted.

(4) Their locations (LOC's) were distributed fairly uniformly over a large geographic region rather than being clustered in a few small spots with large distances between the spots.

Although these characteristics of data point sets used for mapping are highly desirable, much of the data actually available to the MOD effort falls short of these ideals in one or more ways. We emphasize once again that, from the very beginning, the MOD system was conceived as a mechanism to utilize <u>available data</u> as well as possible, relying upon the <u>informed</u> biomedically oriented user to examine output critically and with insight into its limitations.

There are three alternatives to this approach:

(1) Let "well enough" alone.

(2) Set about to collect a great mass of "ideal" medical-environmental data (obviously impractical).

(3) Wait until "ideal" data appears upon the scene.

None of these alternatives was acceptible to us. Our view, as expressed in the Preface, is:

> *Many of the most important problems have the softest information, but we must identify what information there is, and learn its limitations. We must work toward correcting deficiencies in the data base, but even more important, we must develop better methods of using what information is available.*
>
> *We are at the stage of world development where many important judgments must be made in the absence of hard data. If we do not use what data is available, what shall we use?*

4.4 PROBLEM AREAS RELATED TO DATA CHARACTERISTICS

4.4.1 DATA STRUCTURE LIMITATIONS

The method of structuring data which we have described in detail has been used successful by MOD project members to extract disease and environmental data from a variety of sources and, then, to map that data. Furthermore, this structure has provided the basis upon which data extraction forms, data files, and data processing procedures have been designed for the MOD system. But there are limitations to this method of structuring data.

We emphasize once again that the data structure and catalogue of factors are specifically oriented toward mapping. The MOD system was never intended to be a general purpose data-storage-and-retrieval system; rather, it is intended to yield special purpose, disease/environmental maps accompanied by supplementary information. This is why our method of structuring data has developed around the concept of mappable data points. This does not mean that our method of structuring is limited to mapping, but it does imply that modification would probably be necessary if the method were applied to other areas.

Organization of the data structure for computer processing permits data to be entered at any level of generality -- from the very general to the very specific -- and to be retrieved at any level equal to or more general than (but not more specific than) that at which it was originally entered. This means that an individual-patient-record data point can be used in any retrieval, in contrast to (for example) a county-average data point that could never be used to determine the extent of disease in a city within (and smaller than) that county. Generalizing, a computer system can never go beyond the specificity of data within its data pool, but it can always combine (miscible) data into larger groupings. Obviously, then, to be most useful, each data point entered into the MOD system should be as specific and precise as possible to be most useful.

4.4.2 LOCATIONS AND VALUES OF DATA POINTS

Various aspects of locating data points were discussed previously under Output Analysis, but a few additional comments are appropriate to the assignment of values to data points.

The locations of data points input to the MOD system will be stated in terms of political unit names or in terms of longitude-latitude point localities within named political units. (Data grouped for a particular geographical unit will most often be treated as if they all existed at the center-of-gravity of the area of the geographical unit.)

The data point concept which we have discussed is universally applicable to all the data to be put into the system. Conversely, if the data cannot be phrased in terms of discrete data points, it cannot be processed. Therefore, if one wishes to have a disease-environmental map that shows the probable effects of physiographic features such as oceans, deserts, and mountains, he must do so by modifying the particular set of data <u>points</u> used in constructing the map, rather than by modifying the data <u>structure</u>.

For example, there should be no human disease prevalence recorded for the (open) oceans or for (uninhabited) mountain peaks. In order for the system to recognize this situation, the system dictionary must contain a description of all such physiographic features. Once this is done, then such features can be treated as desired. For example, in Fig. 3-38, oceans have been mapped as 0's; in Fig. 3-15, as blank spaces. This is not a unique problem of computer mapping; the human cartographer does this too (Robinson, 1960, p. 160), taking into account many items which are not strictly a part of the data-point set being mapped. The difference between the human cartographer and the computer is that the human cartographer does this intuitively whereas the computer must be given explicit instructions.

In many situations, the assignment of appropriate values to data points has proved extremely difficult, due principally to the incorrect or ambiguous usage of terms relating to disease measures, samples, and extent of diseases in populations. For example, words such as "incidence" and "prevalence" are frequently used interchangeably for several mathematically different ratios or indexes (see Glossary for precise meanings). Careful reading of the data source document sometimes indicates which ratio was meant, but more often, it does not. In this latter case, a professional judgment must be made by the data extractor as to what index was meant. Similar confusion results because sizes of the samples from which published numerical values were calculated are often not given. Suppose the MOD user requests a map that shows infection rate in terms of <u>numbers of infected individuals</u>. To produce this map we must know both the percent infection rate (often given) *and* the number of individuals examined (often not given or, if given, confused with the size of the total population from which the sample was drawn) unless, of course, the <u>number</u> of infected individuals was given in the data source(s).

We hope that development of the MOD system will give further motivation and stimulus to those who generate and report bio-medical data to be precise and rigorously consistent in their use of such terms as we have described, perhaps standardizing definitions in the manner of Dorn, 1957.

4. Data Characteristics

Four types of values for disease data points can be distinguished, and these fall into two major categories:

a. Qualitative, in which alphabetic symbols (or words) denote:
   (1) Occurrence, e.g., "present" or "absent".
   (2) Abundance, e.g., "common" or "abundant".

b. Quantitative, in which numeric symbols (or numbers) denote:
   (3) Absolute numbers, e.g., "10" infected individuals.
   (4) Percentages, e.g., "15" percent infection rate.

Since published reports of disease data include all of the above types of values, the MOD system must contain algorithms suitable for converting any of the types of values to any other. It would be impossible to map together points whose values were "10 cases", "15% infection rate", "common", and "present". All the values must be converted to equivalent values, expressed as only one of the possible types of values. For example, values of "rare" could be converted, either automatically or under user specification, to "2%" or to "5 cases" for compatibility in processing.* Values stated as "N cases", with the smallest and largest sample sizes given as "S_S" and "S_L", can be converted to percentage-type numbers by the formula:

$$\left(\frac{\frac{N}{S_S} + \frac{N}{S_L}}{2} \right) \times 100$$

In certain situations S_S and S_L may be so small that the resulting equivalent percentage-type value for the data point will be artificially high if used alone on a map. Such high values could, perhaps, be suitably marked

---

* The basis for conversion will vary markedly, depending upon the disease-environmental situation. For example: leprosy is considered to be "common" in areas where the prevalence is 2%; a 2% prevalence of dental caries, on the other hand, would warrant the term "uncommon" -- at least in many parts of the world.

and not used in further processing so that the resulting map does not display artificially high disease "peaks".

The problem of values is further complicated by the fact that type of value is related to the factor statement accompanying it -- particularly to the disease measure stated. Somehow, the two must be checked to insure that they are appropriate. For example, in the MOD system, "common", would not be an appropriate value for a data point whose factor stated that the point was for "number of cases existing during specific time period".

In order to map data points at all, each data point must be assigned a single, unique value. This requirement reflects the concept of a map as a mathematical surface (X, Y, Z) in three-dimensional space. By analogy with topographic maps displaying elevation above sea level, it is obvious that a particular geographic point cannot be, simultaneously, 10 feet and 90 feet above sea level. In situations where two specific disease agents were tested for, one found to be present but the other not, there is no conflict with the above statement; two separate data points must be made, one with the appropriate positive value and the other with a zero value; The MOD data structure allows such data to be recorded on a single data extraction form since the data input processing programs can automatically convert these data into one data point with the appropriate positive value and another data point (same LOC) with a zero value (for the agent found to be absent).

The MOD data structure is capable of handling both individual-type (patient/case/clinical) data and group-type (collective/summarized) data. This capability is provided by use of a value, "1" case, plus such LOF's/MOF's as age, sex, occupation, etc. for an individual case, and a value, "23" cases, plus somewhat different LOF's/MOF's to describe the characteristics of a group of (23) cases.

4.4.3 UNRELIABLE DATA

Source documents of medical-environmental data vary greatly in quality. Because of this, an estimate of the reliability of the data is highly desirable, and the MOD system approaches this in two ways.

Reliability of the data source, i.e., trustworthiness, is termed "professional evaluation of data source", and represents a value judgment by the data extractor (or analyst) of the source document's author and his laboratory or institution, as well as the document, per se (experimental design, methodology, etc.). Even though this evaluation will have a highly subjective flavor, it will be much better than nothing. The professional evaluation will probably be limited to a statement of "more reliable", or "less reliable", or "reliability not determined". (A more detailed breakdown of reliability by data extractors has proved unfeasible in our experimental studies.) Some of the factors that would be used by the data extractor in arriving at his decision are:

(1) Are the data published by a highly reputable journal?

(2) Is the report by an author (laboratory, institution) of good reputation?

(3) Is the experimental design good?

(4) Were the experiments done with attention to detail?

(5) Were the conclusions based upon a broad sample (experience)?

(6) Are there contradictions within the report?

(7) Are the stated results completely justified by the observations?

(8) Do the results correspond with those reported from other sources?

(9) Is the study comprehensive?

(10) Do the references cited indicate a thorough background knowledge?

The professional evaluation will be constant for all data points taken from a particular source document. However, since data points within a single

source document can and will vary greatly, we have defined another factor of reliability, termed "computer evaluation of data point". Unlike the former, this evaluation will vary for each data point taken from the document. It will be performed by the computer system itself, according to an algorithm built into the programs that take data into the system. This evaluation concerns the consistency and completeness of each data point. It is based upon specific characteristics found within the data point -- no "judgment" of trustworthiness is involved. For example, one possible algorithm would be to determine if any of the LOF's, on the LOC or VAL of the data point contained "?". If they did, the machine would assign "less reliable" to the point; if they did not, it would assign "more reliable" to the point. Another possible algorithm would allow use of a grading system to represent computer evaluation, each data point to be given a "grade" termed "Computer evaluation number (CEN)." With this method, each MOF would contribute to the total CEN. For example, the MOF, "Time period", could contribute a maximum of 3 points assessed as follows: if the LOF were a year or part of a year, add 2 to the running total; if a range of 2-9 years, add 1 to the total; if a range of 10 or more years, add 0; if no ?'s, add 1; and if ?'s, add 0. Using this method for 12 MOF's, the MOD study team developed a Computer Evaluation Number for leptospirosis data such that the maximum possible CEN = 18. This leptospirosis CEN algorithm was tested with actual data; data points which were judged "good" by all members of the study team tended to have CEN's of about 12, while those judged "bad" tended to have CEN's of about 6.

These two measures of data reliability can (and ordinarily will) be helpful in the process of constructing a map since they would indicate whether or not the data point in question should be retrieved. However, once the data points have been selected, the matter of "reliability" will not enter further into actual construction of the map (but can be an important component of NAR). Data of different reliabilities can be presented as separate output maps if requested. For example, dealing with the same

medical-environmental situation and same geographic area, one map could be based upon only those data points judged "more reliable", a second map based upon only those data points judged "less reliable", and a third map based only upon those data points characterized (by professional evaluation) as "reliability not determined". With the three maps in hand the user could overlay them to obtain a composite map based upon all data points. Then, if he wished, he could "subtract" the "less reliable" and/or the "reliability not determined" data.

Of course there are many limitations of raw data that cannot be resolved in any way short of going back to the individual who made the observations and asking him to clarify or amplify. When the limitations are clearly evident one can guard against misuse, but frequently the limitations are not evident to the data extractor, nor to the data analyst, nor to the computer. The likelihood that such data will be misused is an inherent limitation to any computerized system, including the MOD system. Consider, for example, a report stating that the pH in a particular pond was 5.2. (This information is quite important in the ecology of leptospirosis since the leptospires can survive ((in an infectious state)) quite well in neutral or slightly alkaline water, but quickly die in an acid environment.) The particular water sample may have been taken properly, and the pH measured accurately, yet the data may be seriously misleading. If the pond in question has a very gradually sloping marshy intake side and a dam at the outflow side, the pH could vary markedly. Near the dam (certainly the easiest site from which to take a water sample) the pH might well be 5.0 whereas on the marshy (intake) side it might be as high as 8.

There are many, many other aspects of "unreliability" that are difficult to assess. For example, in many of the developing countries where the major part of the disease data comes from sparse medical centers, the implied geographic distribution of disease may be quite different from its actual geographic distribution since the hospital figures, per se, do not reflect the source of the patient population.

4.4.4 INCOMPLETE DATA

Many highly reliable data are of quite limited use in the MOD system because they are incomplete -- insufficiently characterized as to time or location or factor.

An example of the type of statement frequently found in published papers is: FOUR PERCENT OF CATTLE IN SOUTHERN ILLINOIS HAVE LEPTOSPIROSIS. This apparently straightforward statement leaves many vitally important questions unanswered:

(1) Over what time period were the data collected?

(2) When were the data reported?

(3) If this is a conclusion from a composite of different studies, are we certain that there is no overlapping?

(4) What was the size of the sample(s)?

(5) What are cattle?

- all bovids?
- a limited number of species of bovids?
- a limited number of breeds within one species?
- just cows?
- just mature animals?
- etc.?

(6) What was the nature of the sample(s) of "cattle"?

- sick cattle?
- cattle selected because of the state health department's interest in certain regions?
- cattle selected because of university studies being carried out at specific chosen (e.g. cooperating) farms?

(7) Is it likely that the prevalence was uniform throughout southern Illinois?

(8) What are the precise geographic limits of "southern Illinois"?

(9) What is "leptospirosis"?

disease in terms of:

- -- clinical illness?
- -- detectable antibodies?
- -- recoverable organisms?

(10) What was the inherent accuracy of the diagnostic procedure(s)?

(11) What was the inherent sensitivity of the diagnostic procedure(s)?

(12) How reliable was the laboratory (or laboratories) that performed the analyses?

(13) Were the samples for analyses entirely adequate?

(14) Were the studies which led to this conclusion well planned (i.e., was the experimental design good)?

(15) If this report is a summary/analysis of a collection, is it correct? (i.e., was there an error in transcription or in mathematical manipulation of data?)

(16) Is this report completely honest (i.e., was there intent to mislead)?

Sometimes, in papers of this sort, the (professional) data extractor can infer answers to some of the critical questions, thus increasing the usefulness/applicability of the data. Often, however, answers to the questions simply cannot be gleaned from the report, and data that would otherwise be highly valuable and widely applicable becomes of very limited usefulness.*

4.4.5 CONTRADICTORY AND ERRONEOUS DATA

More often than had been anticipated, we encountered contradictory data in source documents. Sometimes the context indicated which of the two alternatives was correct. Othertimes, however, the only way to resolve the problem was to communicate directly with the author of the paper.

---

* Every effort must be made to see that incompleteness of data is not the fault of the data extractor, and we consider this aspect of the problem in the next section: 5. Data Collection.

A more common -- and more important -- problem comes when different data sources present data that are contradictory. Which of the data are correct is, ultimately, a matter of value judgement.* But the MOD system should (and has been designed to) recognize contradictory data as incompatible. There is only one practical way to recognize such contradictory data; during the synthesis of retrieved data necessary for map construction, the MOD system must check each retrieved data point against all the other retrieved data points. If the LOC and all the LOF's of the factor of the two data points are identical (not just similar), and if the VAL's of the two points are not equivalent, the two points will be deemed contradictory. The system will either let both points stand (to be combined as the user directs, e.g., by averaging) or will call them to the attention of the user (for selection or correction, if desired). Data points which are inconsistent with each other, but not specifically contradictory, cannot be detected directly by the system. Such inconsistencies could be found only by careful examination of the data file by a biomedical professional.

Several other kinds of errors can be detected by the computer system during data input processing and called to the attention of personnel entering the data: for example, data points containing inappropriate values, incorrect LOF's in particular MOF's, and so forth.

4.4.6 SECONDARY DATA POINTS

Published papers nearly always cite work done by other researchers and, as previously explained, the MOD data structure permits construction of secondary data points -- points derived from references quoted by the source document being extracted. It is highly desirable to include such

---

* The MOD system can help in this judgment by providing the "professional evaluation of data source" and the "computer evaluation of data point".

secondary data points for a variety of reasons: the quoted data may be from a personal communication or from an obscure journal, etc., or it may include the explanatory comments of an acknowledged authority, or a critical evaluation of inconsistancies or inadequacies, and the like.

One of the problems with secondary source document concerns duplication, since more than one author may reference the same paper. Duplication might have run to 30 percent in our leptospirosis reprint file without careful control.

Secondary source data creates other problems too. It is often quite incomplete as to data. Furthermore, it represents an incomplete source document reference. These are reasons why secondary source data should be replaced when the primary document is extracted. (Evaluative or explanatory comments need not be discarded.) This requires that bibliographic citation be arranged to allow reference in updating such a point when the primary source is located.

4.4.7 LOCATION TERMINOLOGY

The fact that many relevant papers are written in foreign languages is another data characteristic that contributes difficulty, especially with geographic locations. Place names, e.g., cities and provinces, are usually stated in the native language or in transliterated form. Geographic areas, (e.g., "jungle region" or "coastal region") are also often given in the native language, but these must be translated to become explicit. For example, while extracting data from the province of Bahia, Brazil, reference was made to "Conquista" and to "Litoral" in a context that suggested they were of similar nature, but (as we learned) Conquista is the name of a village, and litoral means coastal. Comparable problems arise when the same locations are reported in different ways, e.g., St. Petersburg/Leningrad, Peking/Peiping, Tokyo/Tokio, etc. Of course, once recognized, many of these can be treated as synonyms. (Others, e.g., St. Petersburg/

Leningrad are not strict synonyms since they relate to different ((historic)) times.) It is not only foreign languages that cause problems of this sort; consider Cape Canaveral/Cape Kennedy.

4.5 TYPES AND CHARACTERISTICS OF DATA SOURCES

Four basically different types of data sources are available to the MOD project, these are:

(1) Published prose summaries (monographs, books, proceedings, journals, technical notes, etc.).

(2) Unpublished prose summaries (progress reports, laboratory reports, letters, and ((oral)) comments, etc.).

(3) Unpublished raw data (field notes, various completed data-collection forms, punched cards, and other items used while preparing, but not included in, published papers).

(4) Published and unpublished maps and photographs.

The scope of these data sources varies immensely. At one extreme lie broad surveys of large regions; at the other extreme are detailed, in-depth studies of small areas or of individual cases.

Ordinarily, the quality of published prose summaries exceeds that of the other types of data, but this quality varies greatly, as anyone familiar with modern scientific literature knows. But quality is a subjective term, and what is good in one context may be poor in another; "good" papers do not necessarily yield good data points for the MOD system. There are several general characteristics of "good" papers that help to explain the reasons for this apparent paradox.

First, most good papers summarize extensive studies. Their purpose is to present (and support) general conclusions concerning the disease situation in a geographical area rather than offer a mass of unprocessed data points. Rarely is all of the data given that the author collected and on

which the paper is based -- and it is usually not possible to infer what the original raw data was. From previous discussion of the MOD data structure, it is apparent that raw data (converted to data points) is the most desirable input to the MOD system. In a sense, the output desired from the MOD system represents a summary made from raw data -- and a good paper accomplishes the same thing (though not in the same form). Another important reason why "good" papers may be poor sources of data for the MOD system is that they often present data in a form that cannot readily be converted to the MOD data point format. And sometimes, although papers state data in a suitable form, they omit one or more of the (six) critical items necessary for the data to be mappable. Or perhaps those data points which can be constructed are too sparse to be handled effectively by current mapping procedures.

A general defect of disease literature is that it lacks environmental data linkable with specific disease data points. In recent years more authors have become aware of the importance of presenting relevant environmental data in their medical papers; hopefully, this trend will continue. But at this time, if only those disease papers presently in our files which had good linkable environmental data were extracted, most of the disease data points necessary to constructing simple disease maps would be lost. This underscores the fact that <u>disease data points</u> need to be extracted even though unaccompanied by good environmental data. Environmental data can be accumulated independently from other sources. Obviously, this is not optimal, but it <u>is</u> a means to bring together data which bears on the same problem/area.

Ideally, data is collected to meet specific requirements, including format. This aspect of data is discussed in detail in Section 5.

Once again we emphasize that, *no matter how efficient the computer manipulation nor how beautifully structured the <u>output information</u>, this information cannot be better than the <u>input data</u>.* Reported variation in a disease-environmental situation may represent <u>actual</u> <u>variation</u> in that

disease-environmental situation, but it may also represent:

- outright fabrication (sometimes politically inspired)
- incorrect generalization based upon inadequate data
- variation in reporting practices
- variation in distribution of medical personnel/facilities
- variation in diagnostic criteria and/or methods
- a very transient influx of persons from a different area (perhaps merely to attend the medical clinic which was collecting the data).
- etc.

This section has been principally concerned with data characteristics. Supplementing this general discussion, a consideration of specific data sources -- narrative as well as map form -- is to be found in the Appendix.

5 Data collection

ABSTRACT - This section considers various data sources and the ways to select, extract and arrange data from those sources. Data extraction procedures are described, and several data extraction forms are reproduced. Then, with the preprocessed data in hand, the methods of entering these into the MOD system are considered.

"Sound generalization can follow only after the determination of precise facts. Data collecting is the indispensable means to synthesis."

Hans Zinsser

5.0 GENERAL CONSIDERATIONS

A very large mass of potentially useful data is available to us -- much more than we can hope to assimilate. We have not forgotten that the primary goal of the MOD project is to develop a computerized disease-mapping system rather than to amass a comprehensive collection of data, but the system must have substance to work upon. The computer processing capability is but one side of the coin; an adequate data file base is the other. And the data that comprises this base must be realistic, reflecting not only actual facts, but what kinds (qualitative and quantitative) of facts are available. Furthermore, the development of more effective methods to acquire, select, extract, and preprocess the raw data is a crucially important part of the MOD system. Thus the need for a significant data collecting effort as an inherent part of the development of the MOD system is clear.

5.1 METHODS OF COLLECTING DATA

There are three basic methods of collecting data:

- Field collection -- This involves direct observation and, in a sense, represents the primary method.
- Literature search -- Literature, here, is used in a broad sense to include various written records -- papers, diagrams, maps, etc. -- published or unpublished. In a sense, this represents a secondary method since the data processor is one step removed from the primary source.
- Combining groups of data collected by others -- In a sense this is a tertiary method since the data processor is two steps removed from the primary source.

The MOD project was never envisioned to have a data generating capacity, and personnel have not been available for field collection, hence this primary method of collecting data was not used (however, in the subsection dealing with data extraction procedures, it will be seen that, through the development of data extraction forms, the MOD system is certainly involved in, and may have an important influence on field collection).

5. *Data Collection*

Literature search was the most appropriate method for collecting data to be used in the MOD system, and most of our data collecting efforts have been expended here. There were two principal reasons for this: first, literature search was the simplest and least expensive method; second, and more important, the MOD system is designed to process available data, i.e., data already in existence. The great bulk of existing data is available from the literature. Obviously, then, we needed to tailor our data collection procedures to fit this most important data source.

Use of large groups of data collected by others was seriously considered, and, if this project had been continued to the point of implementation, it is likely that a portion of the data collection task would have been delegated to an organization such as the Biological Sciences Communication Project (of George Washington University) or the BioSciences Information Service (of Biological Abstracts).

Three broad phases can be distinguished in most data-collecting efforts directed toward literature search. First, source documents containing relevant data must be collected and their contents sufficiently defined (briefly abstracted or summarized) so that they can be filed appropriately. This is the data acquisition phase. Second, the pertinent data contained within the source documents must be extracted (i.e., removed) so that it can be manipulated. This is the data extraction phase. Third, this extracted data must be put in a form (preprocessed) suitable for entry into the computerized information system -- and entered. This is the data entry phase.

The precise mechanism selected for data collecting should depend upon the nature and extent of the data sources, and what resources are available for data collecting. A logical system, which meets the MOD project's (initial) needs, is shown in Figure 5-1. It begins with collection and filing of data source documents by MOD in-house data-acquisition personnel who conduct continuing, comprehensive surveillance of the literature. Data extractors, (who could be medical and/or graduate students, employed

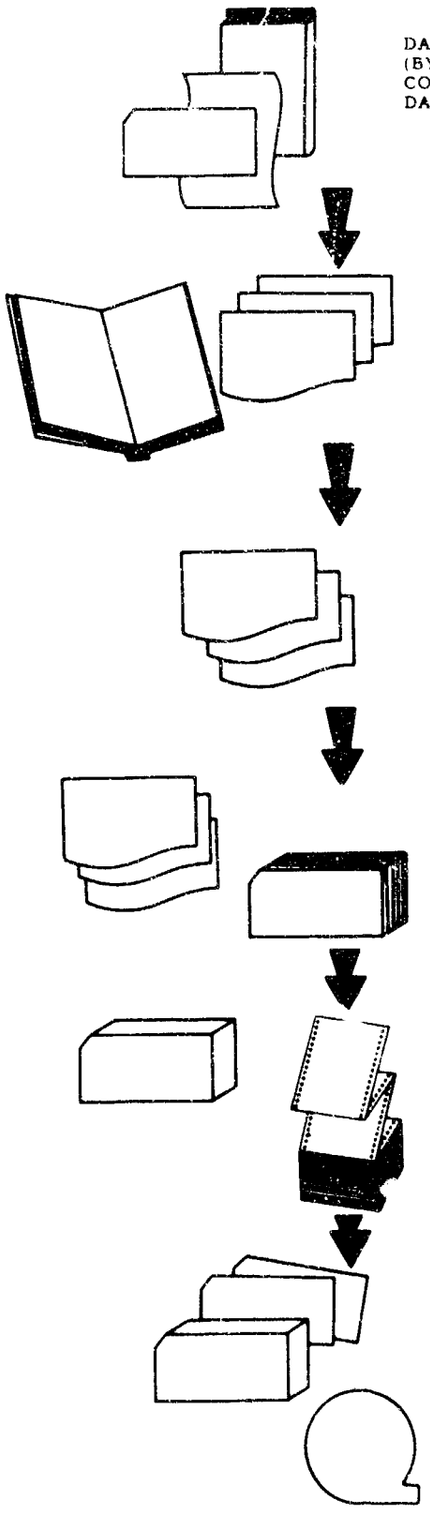


Figure 5-1 A suggested pattern of data collection for the MOD system.

parttime) read the source documents and extract pertinent data from them by filling out data extraction forms that were specifically designed to obtain mappable disease-environmental data. The completed data extraction forms are checked, edited, and corrected by a semi-professional data analyst in coordination with appropriate biomedical professionals who serve as data consultants. Then keypunch operators prepare data input cards, directly from the data extraction forms. Finally, the data analyst has these cards listed, checks the listing, makes appropriate corrections, then enters the cards into the MOD computer system. The organizational chart shown in Section 9 (fig. 9-2, p.9-7), indicates the personnel required for this method of data collection.

5.2 DATA COLLECTION ACTIVITIES

Major efforts at data collection, related particularly to data extraction studies, have concentrated on leptospirosis. (The reasons why leptospirosis was chosen have already been discussed -- p. 1-10). More than 4,500 selected references have been collected, approximately half of which have been abstracted. These source documents were acquired through personal inquiry, bibliographic searches, and a continual surveillance of selected periodicals (professional journals), under the direct supervision of LTC William H. Watson, Jr. (DVM) VC, USAF, Chief of the Geographic Zoonoses Branch of the Division of Geographic Pathology of the Armed Forces Institute of Pathology (AFIP). (For more detailed information about data sources see the Appendix.) These leptospirosis data served as the basis for setting up the MOD source document storage-and-retrieval system and greatly influenced our thoughts in designing the MOD data structure, the factor catalog, the data extraction forms, and the computerized data files.

Data collection and extraction, on the scale that we have discussed, is a long, slow, tedious process. While this aspect of the program was developing, we sought a readily limited "package" of data which would allow us to begin work on defining and assembling data points, and producing maps

from these data points. Such a package was obtained from data published by Malek (in Studies of Disease Ecology, edited by May, J.M., 1961), relating to schistosomiasis in South America and Africa. Although these data are limited in their extent and are no longer current, they have been very helpful to us. Use of these data has provided valuable insight into both cartographic procedures and problems in data structuring.

As the MOD system design progressed, it became necessary to test further some of the computer-mapping programs under consideration. Data required to make these tests had to be readily available in a simple standardized (preferably tabular) format and, in addition, it had to be rather uniformly and densely distributed over a relatively large geographic area. The National Communicable Disease Center supplied us with several sets of unpublished data on rabies in the eastern U. S. that entirely satisfied these requirements. Unfortunately, support of the MOD project terminated before full use could be made of this material, although it did serve its immediate purpose, and several maps have been produced from it. For the same reason, i.e., termination of support, we have not taken full advantage (by far) of the extensive leptospirosis data that we have acquired during the past two and a half years. This collection has fulfilled a very important purpose, however, and will serve effectively as a major component of the data file base if and when the MOD system is implemented.

5.3 DATA EXTRACTION PROCEDURES

Once the data sources have been collected and organized, the task of extracting the relevant data from the sources begins. When the MOD project began, we did not realize the extent to which this would pose unusual problems. However, the further our work progressed, the more apparent it became that data extraction/preprocessing presented problems that were far more complex than could have been anticipated in the beginning. Some of these problems have been encountered by other groups, but more often avoided rather than resolved. Some of the important problems in this area have not

been encountered by other groups so far as we can determine, probably because no one else has attempted to develop the kind of system which MOD represents.

Our basic approach to solving these data-extraction problems has been through a group effort, involving both data-processing and data-collecting personnel. Repeated attempts to extract and put into consistent form the data on disease and environmental factors contained in selected representative data sources were carried out, until, finally, the extracted data were in a form acceptable to the data processors as well as the data collectors/analysts. As a result of this extensive trial and error method, general requirements for data content/format were formulated, and this has been one of our most important accomplishments (largely due to the efforts of Dr. Cuffey).

Our first major problem was that no generic terms existed which encompassed disease-environmental data. Thus it became necessary to construct a general data-analysis vocabulary before we could communicate effectively in relation to the disease-environmental data which we were attempting to extract. This data-analysis vocabulary includes definitions for and discussion of the interrelationships among such vitally important terms as "factor", "common elements", "value", "data point", "map", and "narrative".

The second major problem was to specify precisely what items of disease-environmental information were pertinent to our major objective: the production of disease distribution maps. This led to the development of a catalogue of disease-environmental factors that could be used by the MOD computer system in producing disease-environmental factor distribution maps. These two aspects of the data problem are discussed in the preceeding section and (at least) the material contained in pages 4-4 through -13 should be read before considering the details of data extraction.

We have found that many of the data available for processing are incomplete in one way or another and, often, professional judgement/interpretation (sometimes extrapolation) must be carried out if the data are to be

usable. Narrative print-out, to accompany the computer maps, will note these interpretations, and source document numbers will be available upon request should the user wish to consult the data source. Some of the data will be of very limited use because essential factors (which must have been known to the author) simply aren't recorded. These problems are numerous and serious (as discussed in the preceeding section), but we must do the best we can with the information available.

Extraction form design is the key to successful data extraction. The development of extraction forms has proved to be exceptionally difficult because of the extremely varying content of the data sources, coupled with the requirement that the data must relate to a consistent, geographically-oriented format.

Initially, efforts at designing data extraction forms vacillated between the use of free-format and fixed-format styles. The disadvantages of each were usually more apparent than the advantages. Our early experience led us to discount the use of fixed-format data within the computer data files, and this bias carried over to design of the forms. This was primarily because whenever fixed-format data is recorded, the resulting system becomes limited.

We have come to the view that the freer the format of extraction, the less clearly evident what data is desired. In addition, the freer the format the longer it takes to extract the data from a paper, and the more difficult it becoms to reformat the data for computer entry. Even more important, there will be a greater loss of potentially useful data. A more rigid format is better, primarily because, even though the literature is extremely varied, a biomedical person can translate effectively many of the variations into correspondences using a few standard items that will ultimately be easier to query. Also (as a corollary of the statements about freer format), the shorter and simpler the data form, the more data points are likely to be extracted, thereby ensuring a more useful output.

5. *Data Collection*

Fixing the format of a data form in no way limits the computer processing. The data form is designed solely to guide the extractor as to what data is desired, and to insure uniform data extraction. For example, one data form can be designed for leptospirosis, another for schistosomiasis while still another can be used for certain environmental factors. All the data can go into the same data file for computer processing -- and all the data (assuming that has been properly formatted) is miscible (and, obviously, usable in an almost infinite variety of contexts).

Data forms were designed with careful consideration of the MOD data structure and catalog of factors, and in consultation with various biomedical and data-processing professionals. The forms were then tested (actually used with source documents), modified as necessary, retested, etc., etc. until they appeared to be both effective and efficient. By requiring (some) items to be recorded in a fixed format rather than "natural language", simple codes were utilized, on the forms, to facilitate keypunching. Furthermore, these codes were constructed so as to allow the computer system to perform certain kinds of error checking of the input data. The latest leptospirosis data extraction forms are given (Fig. 5-2 and 5-3), also forms that were used to record schistosomiasis (Fig. 5-4) and rabies (Fig. 5-5) data. One example of an environmental data extraction form, shown in Fig. 5-6, was used to record data relating to the geographic distribution of certain small mammals (particularly important to understanding the epidemiology of leptospirosis). Another type of environmental data collection form, shown in Fig. 5-7, was used in compiling a file of published maps dealing with environmental factors of southeast Asia.

Our tentative scheme for handling the collected, selected data is a three-stage process:

(1) Data extractors (necessarily with biomedical background since value judgements are required) will fill in relatively simple data-extraction forms. These forms will be submitted to a

—— continued page 5 - 17

MOI DATA EXTRACTION FORM — FOR MINIMAL LEPTOSPIROSIS DATA

(SEE PAGE [illegible] OF FACTORS FOR EXPLANATION OF MOF's/LOF's)

Security classification of Data: [ ] (not to be keypunched)

Data Point Number: year — month — day — Cartographer's initials — point number

(LOC) continent/ocean ; country/colony/dependency ; state/province ; county/parish ; very small unit

Geographic Location

c/w longitude — degrees — minutes — tenths of degree

n/s latitude — degrees — minutes — tenths of degree

point locality

(VAL) _____ Value for Data Point

(DMS) Disease Measure [greatly influences value of data point, which is recorded above]

- 321 ☐ Occurrence
- 225 ☐ Abundance
- 327 ☐ Point prevalence
- 426 ☐ Period prevalence
- 525 ☐ Incidence
- 624 ☐ Mortality
- 723 ☐ Standardized mortality ratio
- 822 ☐ Number of cases existing at specific point in time
- 921 ☐ Number of cases existing at any time during specific time interval
- 1025 ☐ Number of cases beginning during specific time interval
- 1124 ☐ Number of deaths during specific time interval
- --2- ☐ Other (specify: _____)

PRECISE IDENTITY — DOMESTICATION (Domesticated / Wild)

Animal Infected: (AIP) _____ (AID) 196, ☐ 295, ☐

(GKD) 262 ☒ Leptospirosis — General Kind of Disease

(SDA) Specific Disease Agent

- 471, ☐ Leptospira, species/serotype not specified or undifferentiated
- 570, ☐ Leptospira, all species/serotypes
- -7-, ☐ Leptospira, species/serotypes indicated present (specify: _____)
- -7-, _____
- -7-, _____
- -7-, _____ (etc.)
- -7-, ☐ Leptospira, species/serotypes indicated absent (specify: _____)
- -7-, _____
- -7-, _____
- -7-, _____ (etc.)

(ESD)

- 118, ☐ Endemic/enzootic
- 217, ☐ Hyperendemic/hyperenzootic
- 316, ☐ Sporadic
- 415, ☐ Epidemic/epizootic
- 514, ☐ Pandemic/panzootic

Epidemiologic State of Disease within Population

(TIM) from: year — month — day ; to: year — month — day — Time Period for Which Data Applies

(MDG) Method of Diagnosis

- 125, ☐ Clinical observation
- 224, ☐ Isolation, other/unspecified
- 323, ☐ Isolation from water
- 422, ☐ Isolation from soil
- 521, ☐ Isolation from wine
- 620, ☐ Isolation from blood
- 723, ☐ Isolation from tissue
- 888, ☐ Serologic test
- 987, ☐ Xeradiagnosis
- 1051, ☐ Biopsy
- 1150, ☐ Autopsy
- --8-, ☐ other (specify: _____)

(UDS) year — month — day — collector initials — sample number — Unique Designator for Combination of Smallest and Largest Samples Involved

	FOR SMALLEST SAMPLE INVOLVED	FOR LARGEST SAMPLE INVOLVED
Size:	(SSZ) _____	(LSZ) _____
Age:	(SSA) _____	(LSA) _____
Sex:	(SSX) 285, ☐ Male 284, ☐ Female	(LSX) 196, ☐ Male 295, ☐ Female
Racial/Ethnic Groups:	(SRE) _____	(LRE) _____
Occupational Group:	(SOG) _____	(LOG) _____
Basis for Sampling:	(SBS) _____	(LBS) _____

(PSD) author(s) ; date ; journal/book/report ; volume ; number(s) — Primary Source Document

(SSD) author(s) ; date ; journal/book/report ; volume ; number(s) — Secondary Source Document

(PES) 274 ☐ More reliable 273 ☐ Less reliable 372 ☐ Reliability not assessed — Professional Evaluation of Data Source

(CEI) _____ [every value to be calculated automatically by MOI system] — Computer Evaluation of Data Point

Other LOF's/MOF's and NAR: (write on back of form)

Figure 5-2

MOD DATA EXTRACTION FORM FOR COMPREHENSIVE LEPTOSPIROSIS DATA

(SEE MOD CATALOG OF FACTORS FOR EXPLANATIONS OF MOF's/LOF's)

Security Classification of Data [not to be keypunched]: ☐

Data Point Number

year ___ month ___ day ___ extractor's initials ___ point number ___

(LOC) continent/ocean ___ , country/colony/dependency ___ , state/province ___ , county/parish ___ , very small unit ___ , point locality ___

Geographic Location

E/W longitude ___ degrees ___ minutes ___ tenths of degree

N/S latitude ___ degrees ___ minutes ___ tenths of degree

(VAL) ___

Value for Data Point

(DMS)
- 123 ☐ Occurrence
- 228 ☐ Abundance
- 327 ☐ Point prevalence
- 426 ☐ Period prevalence
- 525 ☐ Incidence
- 624 ☐ Mortality
- 723 ☐ Standardized mortality ratio
- 822 ☐ Number of cases existing at specific point in time
- 921 ☐ Number of cases existing at any time during specific time interval
- 1025 ☐ Number of cases beginning during specific time interval
- 1124 ☐ Number of deaths during specific time interval
- --2- ☐ Other (specify: ___)

Disease Measure [greatly influences value of data point which is recorded above]

	PRECISE IDENTITY	DOMESTICATION (AID) Domesticated	Wild
Animal Infected: (AIP)	___	(AID) 196, ☐	255, ☐
Primary Host: (PHP)	___	(PHD) 118, ☐	217, ☐
Intermediate Host: (IHP)	___	(IHD) 130, ☐	239, ☐
Reservoir: (RSP)	___	(RSD) 152, ☐	251, ☐
Carrier: (CRP)	___	(CRD) 174, ☐	273, ☐
Vector: (VCP)	___	(VCD) 196, ☐	255, ☐

(GKD)
- 262, ☐ Leptospirosis
- -6-, ☐ Other (synonymous with leptospirosis; specify: ___)

General Kind of Disease

(SDA)
- 471, ☐ Leptospira, species/serotype not specified or undifferentiated
- 570, ☐ Leptospira, all species/serotypes
- -7-, ☐ Leptospira, species/serotypes indicated present (specify: ___)
- -7-, ___
- -7-, ___
- -7-, ___
- [etc.]
- -7-, ☐ Leptospira, species/serotypes indicated absent (specify: ___)
- -7-, ___
- -7-, ___
- -7-, ___
- [etc.]

Specific Disease Agent

(ESD)
- 113, ☐ Endemic/enzootic
- 217, ☐ Hyperendemic/hyperenzootic
- 316, ☐ Sporadic
- 415, ☐ Epidemic/epizootic
- 514, ☐ Pandemic/panzootic

Epidemiologic State of Disease within Population

(TIM) from: year ___ month ___ day ___ to: year ___ month ___ day ___

Time Period for which Data Applies

(SEA)
- 107, ☐ Winter
- 206, ☐ Spring
- 305, ☐ Summer
- 404, ☐ Fall
- -0-, ☐ Other (specify: ___)

Season for Which Data Applies

NOT REPRODUCIBLE

Figure 5-3 Parts A and B. MOD data form for extracting more than "minimal" leptospirosis data for mapping.

(See next page for B.)

COMPREHENSIVE LEPTOSPIROSIS DATA — PAGE 2

(MDG) 185, ☐ Clinical observation
284, ☐ Isolation, other/unspecified
383, ☐ Isolation from water
482, ☐ Isolation from soil
581, ☐ Isolation from urine
630, ☐ Isolation from blood
783, ☐ Isolation from tissue
882, ☐ Serologic tests
957, ☐ Xerodiagnosis
1081, ☐ Biopsy
1180, ☐ Autopsy
--8--, ☐ Other (specify: _____)
Method of Diagnosis

(MFD) 196, ☐ Military hospital/clinic
295, ☐ University/academic hospital/clinic
394, ☐ Large/urban hospital/clinic
493, ☐ Small/rural hospital/clinic
592, ☐ Individual physician
691, ☐ Nurse/paramedical person
790, ☐ Folk/witch doctor
889, ☐ None
--9--, ☐ Other (specify: _____)
Medical Facilities Involved in Diagnosis

(UDS) _____ year _____ month _____ day _____ extraordinarily _____ sample number
Unique Designator for Combination of Smallest and Largest Samples Involved

	FOR SMALLEST SAMPLE INVOLVED	FOR LARGEST SAMPLE INVOLVED
Size:	(SSZ) _____	(LSZ) _____
Age:	(SSA) _____	(LSA) _____
Sex:	(SSX) 185, ☐ Male 284, ☐ Female	(LSX) 196, ☐ Male 295, ☐ Female
Racial/Ethnic Group:	(SRE) _____	(LRE) _____
Occupational Group:	(SØG) _____	(LØG) _____
Basis for Sampling:	(SBS) _____	(LBS) _____

(MTR) 107, ☐ Direct contact with living infected animals
206, ☐ Direct contact with dead tissue or blood
305, ☐ Direct contact with excreta (incl. urine)
404, ☐ Indirect occupational contact with water
503, ☐ Indirect recreational contact with water
602, ☐ Indirect domestic contact with water
701, ☐ Indirect occupational contact with soil
800, ☐ Indirect recreational contact with soil
909, ☐ Indirect domestic contact with soil
1003, ☐ Bite of carrier or vector
--0--, ☐ Other (specify: _____)
Method of Transmission to Animal Infected

(IMU) 130, ☐ Susceptible/not immune
239, ☐ Naturally immune
338, ☐ Artificially immunized
Prior Immunity of Animals Infected in This Outbreak

(KØR) 129 ☐ Isolated (1) case 228 ☐ Small group (2-29) of cases 327 ☐ Large group (30+) of cases
Kind of Outbreak Reported

(FØP) 170 ☐ No outbreaks previously reported
239 ☐ Outbreaks rare/occasional/seldom
338 ☐ Outbreaks common/frequent
437 ☐ Outbreaks very common/very frequent
Frequency of Outbreaks Preceding Outbreak Reported

(DØR) _____ [days]
Duration of Outbreak Reported

SEVERITY OF DISEASE IN THIS OUTBREAK REPORTED

		Fatal	Severe clinical	Moderate clinical	Mild clinical	Asymptomatic/Subclinical
Average:	(AVS)	152 ☐	251 ☐	350 ☐	459 ☐	558 ☐
Minimum:	(MNS)	163 ☐	262 ☐	361 ☐	460 ☐	569 ☐
Maximum:	(MXS)	174 ☐	273 ☐	372 ☐	471 ☐	570 ☐

(LDØ) 185 ☐ Always fatal
284 ☐ Often fatal
383 ☐ Seldom fatal
482 ☐ Rarely fatal
581 ☐ Never fatal
Lethality of Disease in This Outbreak

(AVC) 196 ☐ Acute
295 ☐ Subacute
394 ☐ Subchronic
493 ☐ Chronic
Average Course of Disease in This Outbreak

DURATION OF CASES IN THIS OUTBREAK

Average: (AVD) _____ [days]
Minimum: (MND) _____ [days]
Maximum: (MXD) _____ [days]

(MFT) 141, ☐ Military hospital/clinic
240, ☐ University/academic hospital/clinic
349, ☐ Large/urban hospital/clinic
448, ☐ Small/rural hospital/clinic
547, ☐ Individual physician
646, ☐ Nurse/paramedical person
745, ☐ Folk/witch doctor
844, ☐ None
--4--, ☐ Other (specify: _____)
Medical Facilities Involved in Treatment during This Outbreak

(PSD) _____ author(s), _____ date,
_____ journal/book/report, _____ volume, _____ number, _____ page(s)
Primary Source Document

(SSD) _____ author(s), _____ date,
_____ journal/book/report, _____ volume, _____ number, _____ page(s)
Secondary Source Document

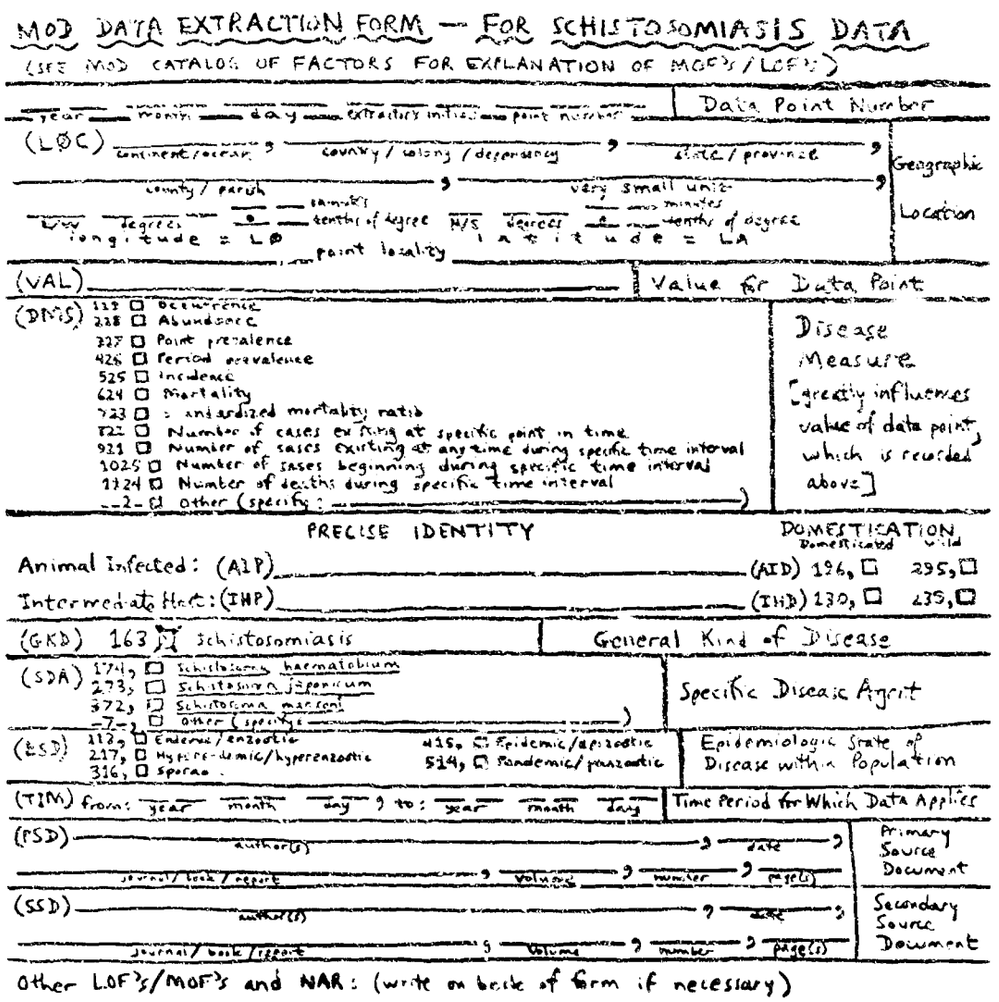
(RES) 174 ☐ More reliable 273 ☐ Less reliable 372 ☐ Reliability not assessed
Professional Evaluation of Data Source

(CEL) _____ [eventually to be calculated automatically by MOD system]
Computer Evaluation of Data Point

Other LOF's/MOF's and NAR: (write on back of form)

NOT REPRODUCIBLE

Figure 5-3 -- continued



MOD DATA EXTRACTION FORM — FOR SCHISTOSOMIASIS DATA

(SEE MOD CATALOG OF FACTORS FOR EXPLANATION OF MOF's / LOF's)

year month day extractor's initials point number — Data Point Number

(LOC) continent/ocean ; country / colony / dependency ; state / province ; county / parish ; very small unit ; Geographic Location

Long. degrees ° minutes ' tenths of degree N/S degrees ° minutes ' tenths of degree Lat.

longitude = LO point locality latitude = LA

(VAL) — Value for Data Point

(DMS) Disease Measure [greatly influences value of data point, which is recorded above]

- 112 ☐ Occurrence
- 238 ☐ Abundance
- 327 ☐ Point prevalence
- 426 ☐ Period prevalence
- 525 ☐ Incidence
- 624 ☐ Mortality
- 723 ☐ standardized mortality ratio
- 822 ☐ Number of cases existing at specific point in time
- 921 ☐ Number of cases existing at any time during specific time interval
- 1020 ☐ Number of cases beginning during specific time interval
- 1224 ☐ Number of deaths during specific time interval
- --2- ☐ Other (specify:)

PRECISE IDENTITY — DOMESTICATION (Domesticated / Wild)

Animal Infected: (AIP) — (AID) 196, ☐ 295, ☐

Intermediate Host: (IHP) — (IHD) 130, ☐ 439, ☐

(GKD) 163 ☒ Schistosomiasis — General Kind of Disease

(SDA) Specific Disease Agent

- 1741 ☐ Schistosoma haematobium
- 2732 ☐ Schistosoma japonicum
- 3722 ☒ Schistosoma mansoni
- --7-- ☐ Other (specify:)

(ESD) Epidemiologic State of Disease within Population

- 112, ☐ Endemic/enzootic
- 217, ☐ Hyperendemic/hyperenzootic
- 316, ☐ Sporadic
- 415, ☐ Epidemic/epizootic
- 514, ☐ Pandemic/panzootic

(TIM) from: year month day ; to: year month day — Time Period for Which Data Applies

(PSD) author(s) ; date ; journal / book / report ; volume ; number ; page(s) — Primary Source Document

(SSD) author(s) ; date ; journal / book / report ; volume ; number ; page(s) — Secondary Source Document

Other LOF's/MOF's and NAR: (write on back of form if necessary)

NOT REPRODUCIBLE

Figure 5-4 MOD data form for extracting standard schistosomiasis test data (given earlier in Fig. 3-1).

MOD DATA EXTRACTION FORM —
FOR UNPUBLISHED RABIES DATA
(SEE MOD CATALOG OF FACTORS FOR EXPLANATION OF MOFs/LOFs)

year — month — day — extractor's initials — point number	Data Point Number
(LØC) N AM, UNITED STATES, state, county, — degrees minutes longitude + degrees minutes latitude coordinates of center-of-area of particular county	Geographic Location
(VAL) number	Value for Data Point
(DMS) 921 Number of ca existing at any time during specific time interval	Disease Measure
(PIP)	Precise Identity of Animal Infected
(GKD) 361 Rabies	General Kind of Disease
(TIM) year	Time Period for Which Data Applies
(PSD) RABIES UNIT OF NATL COMMUNIC DIS CTR, year, ANIMAL RABIES FORMS — PHS 4.198 CDC.	Primary Source Document

Figure 5-5 MOD data form for extracting rabies data (see Fig. 3-44).

MOD DATA EXTRACTION FORM
FOR SMALL-MAMMAL DISTRIBUTION DATA
(SEE MOD CATALOG OF FACTORS FOR EXPLANATION OF MOF's/LOF's)

year month day extrallary initials point number — Data Point Number

(LOC) continent/ocean ; country/colony/dependency ; state/province ; county/parish ; very small unit — Geographic Location

E/W degrees longitude — minutes / tenths of degree — point locality — N/S degrees latitude — minutes / tenths of degree

(VAL)
- DP ☐ Definitely present
- PP ☐ Probably present
- PS ☐ Possibly present/possibly absent
- PA ☐ Probably absent
- DA ☐ Definitely absent
- NL ☐ Not looked for

Value for Data Point

(DMZ) 125 Occurrence — Distribution Measure

(SMP) _____ — Precise Identity of Small Mammal Considered

(TIM) from: year month day ; to: year month day — Time Period for Which Data Applies

(PSD) author(s) ; date ; journal/book/report ; volume ; number ; page(s) — Primary Source Document

(SSD) author(s) ; date ; journal/book/report ; volume ; number ; page(s) — Secondary Source Document

other LOF's/MOF's and NAR: (write on back of form if necessary)

Figure 5-6 MOD data extraction form for d[illegible]g to biogeographic distribution of small mammals.

NOT REPRODUCIBLE

ENVIRONMENTAL-FACTOR MAP INFORMATION FORM (mark all appropriate blanks)

1. Scope of map: ☐ world (entire) ☐ SE Asia (entire) ☐ Mid-Western U.S.
☐ World (major sections; continent, ocean; specify: _____) ☐ Thailand ☐ Illinois (entire)
☐ Malaya ☐ southern Illinois (including quadri-county area)
☐ other (specify: _____)

2. EXACT title of map: (i.e., name of environmental factor mapped)

3. Breakdown of values of environmental factor mapped — XEROX LEGEND OF MAP AND STAPLE TO THIS SHEET

4. Means of representing data on map:
☐ dot-type symbols ☐ shading/patterns (black-white) ☐ contour-type lines
☐ alphabetic/numeric symbols ☐ shading/patterns (colors) ☐ other (specify: _____)

5. Projection used:
☐ equirectangular (plane-chart)
☐ Miller cylindrical
☐ mercator
☐ homalosine
☐ other (specify: _____)
(also note whether:
longitude meridians are ☐ straight or ☐ curved lines
latitude parallels are ☐ straight or ☐ curved lines
projection is ☐ interrupted, ☐ interrupted and condensed, ☐ condensed, or ☐ non-interrupted and non-condensed)

6. Scale of map: 1/_____

7. Dimensions of map: _____ cm. / _____ in. X _____ cm. / _____ in.
(from unit mentioned above) (width) (height)

8. Date of publication of map: _____

9. Date(s) of data mapped: _____ (earliest) — _____ (latest)

10. Data compiled by (if different from Item 11): _____

11. Bibliographic reference for map: (book — author, date, title of book; publisher and city; page
journal — author, date, title of article: name of journal, volume: (number): page)

12. Call number of source containing map: _____

13. Call number of map (if different from Item 12): _____

14. Physical location of map (or source containing map):
☐ Univ. of Illinois main library
☐ Univ. of Illinois map library
☐ faculty member's personal library (specify whose: _____)
☐ AFIP Asia library
☐ AFIP Geographic Path.-Geographic Zoon. library
☐ staff member's personal library (specify whose: _____)
☐ other location (specify: _____)

15. Additional remarks (use reverse side if necessary):

16. This form was completed on: by:

Date: _____ Name: _____

Figure 5-7 MOD data collection form used in compiling file of published environmental-factor maps of southeast Asia.

data-analyst(s), who, with the help of data-consultants, as necessary, will check (edit) the forms

(2) The data analyst will transcribe the data from the extraction form to a more rigidly formatted (intermediate) form.

(3) The intermediate form will be converted to punched cards for input into the computer system.

Elaborating upon phase 1 (above), the data extractor first obtains a suitable source document (ordinarily from the MOD data source storage-and-retrieval files). Then he skims the document, selects blocks of potential data points, partially fills in several forms (with items common to several points), and, using a duplicating machine (e.g., Xerox), makes a number of copies of each partially completed form. Using Xeroxed forms and unicolored pencils can save much time because of the great number of duplicate entries required to extract many data points from one source document. An alternative possibility would be to enter all the common items for each document on a single master sheet, and to provide a reference by which the data points' forms could be connected to this master sheet. Since most documents contain several groups of data points for which several master sheets would have to be constructed, we have found it more efficient to make duplicate copies of partially completed forms during the data extraction.

Next, the data extractor carefully goes back through the source document and completes one data form for each data point. He then adds "other LOF's/MOF's" (and NAR) as necessary to ensure that each data point is adequately and completely defined. Finally, he talks with appropriate data consultants to resolve any remaining questions about particular data points.

Based upon a relatively small but diverse and representative sample of data source documents, we have found that approximately two papers, each averaging 10 pages, can be extracted in depth by one data extractor in a day. With training and experience data extractors could certainly work faster than this, but not a great deal faster so long as they extracted

(virtually) all of the significant data. Obviously, there are far too many disease-environmental factors described in the published literature to extract all, or even most, of them. The most rational approach is to be selective: to extract in depth only for highly relevant factors, and (unless there is particular reason not to) restrict attention to those items which are necessary to construct data points (as described in Section 4). Data points with many qualifying LOF's/MOF's may be used so seldom in retrieval and mapping operations that their value is not worth their cost. These two methods of approach, coupled with careful selection of the source documents, will help with the data volume dilemma, and still provide sufficient kinds and numbers of data points to make computer processing worthwhile. We have mentioned before that a large volume of environmental data is already published in map form and that computerized maps of the type we are discussing can be made to fit with (e.g., overlay) a published map, obviating the need to extract and process that data which has already been mapped. The use of secondary source documents (compiled data) and data presented in tabular form would also materially reduce extraction requirements.

In the discussion of time and effort involved in data extraction we have considered primarily the average time required to extract "representative" source documents, but the source documents vary enormously. Some short papers yielded as many as 50 data points; others, of comparable size, yielded only one. Furthermore, the structure and the language of source documents can make it difficult (and time consuming) or easy (and relatively quick) to extract the data necessary to produce data points.

Based on our experiences, there will be considerable personnel problems among data extractors. Extraction efforts are very demanding and fatigue develops far out of proportion (seemingly) to the work actually accomplished. Furthermore, we observed lack of day to day consistency in the type of data actually recorded on the data extraction forms. The task is very boring and it is easy for the extractor to become distracted. In

our experiences several people, each working part-time, were better than one working full-time since this reduced boredome and the frequency of distraction. Some consistency is lost, but,on different days, the output of a single extractor was as inconsistent as that from several different extractors. It seems likely that but a small proportion of persons will be found to have the psychologic and intellectual qualities necessary to become excellent data extractors.

Many problems appeared during our attempts to develop data extraction techniques. Those that relate to data characteristics, such as incompleteness and unreliability of data points, assignment of values, accurate specification of disease measures and samples involved, and precise statement of the geographic locations of data points, have already been discussed. The great variation in quality of data sources, as well as in the numbers of data points extractable from single source documents, have also been considered. Perhaps the most important difficulty lies in the fact that much of the material in a typical narrative paper is not mappable. Moreover, that material which is mappable may require the extractor to read large sections of the paper and then piece together -- using professional judgment -- the few disconnected fragments of critical data, floating in a sea of words that contribute no data but yet, somehow, seem necessary for the paper to be understandable. Extracting data is clearly <u>not</u> a simple matter; there is no automatic procedure by which narrative sentences can be converted into computer-mappable data points. Putting it another way, "data processing" (at this level) becomes heavily involved with "communication science" and general semantics.

* * *

<u>In summary</u> -- our experiences with data-management in the context of this comprehensive program have exposed complexities of a degree that we could not have anticipated. It has become clearly evident that *the most critical factor limiting meaningful computer output of the MOD system is the content/format of input data. The sources of the data are readily*

available, but there are major difficulties in extracting/formatting these data. These problems relate to:

(1) Highly varying source document content (requiring development of a data-analysis vocabulary and a factor catalogue to establish common denominators).

(2) Highly varying reliability of raw data (requiring a system for defining reliability and, on occasion, validating data).

(3) Necessity for continual changes in and additions to the data base file (making unusual requirements for editing and updating).

(4) Lack of a generic vocabulary encompassing medical-environmental situations (related to item #1).

(5) Inherent complexities in the data which make it difficult to specify a standardized procedure(s) for the extraction, editing, structuring, and storing of the data prior to computer input.

(6) Data file design problems due to complexities of the data in general, its great volume and the large number of interrelationships among the specific data and among descriptions associated with vocabulary/definitions after computer input.

But turning to a more positive view, our efforts have resolved (in large measure) most of these problems. The evidence for this is in the form of computer produced medical-environmental maps, maps derived from data that were collected, extracted, and input in the manner described in this section.

5.4 DATA INPUT OPERATIONS

The last phase of the MOD data collection process involves transferring the data contained on the data extraction forms into the MOD computer system data files. The basic procedures for this are illustrated by an actual example as shown in Fig. 5-8.

The data extraction forms filled out by the data extractors are first examined, edited, and corrected, if necessary, by a data analyst. This person, functioning as an intermediary between the data extraction personnel

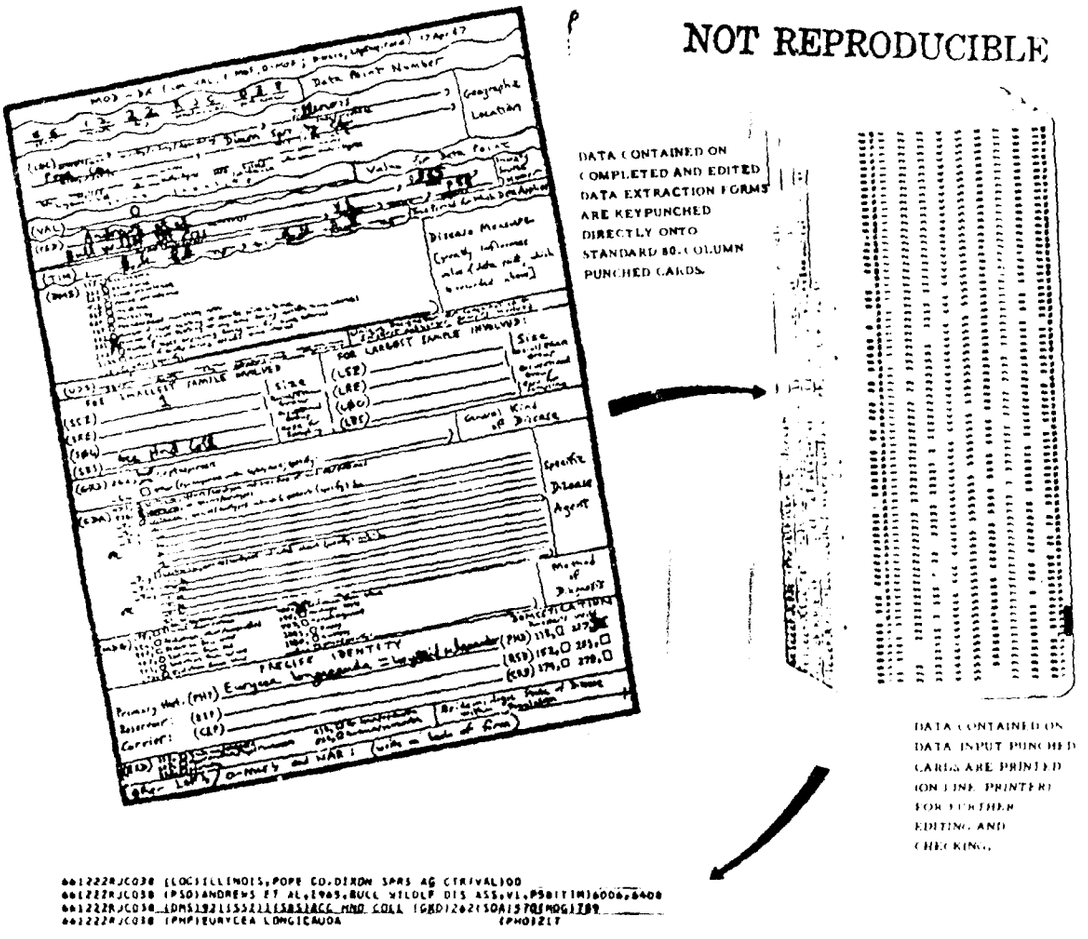


Figure 5-8 Basic procedures followed in entering data into the MOD computer system, using an actual leptospirosis data point as an example.

and the computer system, must be familiar with both biomedical and automatic data processing matters so that he can detect (and correct) erroneous or inappropriate data point values, inconsistencies and errors in data format, unallowable LOF's in particular MOF's, omission of critical items in the data point records, and the like. Questions which the data analyst cannot resolve will be answered by the data consultant.

The data analyst then gives the edited data extraction forms to keypunch operators who punch the data directly into standard 80 - column punch cards. These cards have been chosen as the input medium for the MOD system (as discussed in the section on computer system requirements) primarily because of the flexibility which they provide in all phases of data processing.

Compromise is inevitable in selecting the <u>form of the input</u>. A natural or problem oriented language is easier for the data processor to use whereas a fixed-form input format is easier for the computer to handle. A proper compromise is one in which the kind of language input format developed best suits the <u>total</u> procedure. Arriving at this proper compromise represents a critical step since an inappropriate selection would lead to much delay and costly duplication of effort. Before we took this critical step a number of trials were conducted, and these involved use of fixed-field and variable-length records, numerical codes, alphanumeric (mnemonic) codes, and essentially natural language-type statements. We concluded that information appearing on the data input cards should be in a fixed format -- partly fixed card-column, but mostly fixed order and punctuation. (The data input card formats are discussed in detail later in connection with data processing operations.) Because relatively little data has been key punched according to the full data point format developed for the MOD system, we cannot give precise estimates of the effort required to process large numbers of data extraction forms.

The next step in the data input process is for the data analyst to input the punched cards to the MOD computer system. The system reads the

cards, performs the kinds of error checking for which it is programmed, prints out a listing showing all the data cards and also indicates the errors found.

The data analyst carefully examines the listing for possible errors of the kind which the system cannot detect. He corrects the errors which he finds, and those which the system noted, by repunching appropriate data input cards.

Finally, the data analyst puts the corrected deck of punched cards back into the computer system. The system reads the cards and stores their contents internally in its various files.

At this point, the data collecting tasks are completed and the computer system is prepared, so far as its data file base is concerned, to manipulate data in response to query.

* * *

We emphasize once again the importance and difficulty of the research effort that has been necessary to process the raw data <u>before</u> they reach the computer. Without adequate and properly formatted input, significant output is impossible. The old term, GIGO, expresses the situation well: garbage in/garbage out

6 Computer system requirements

ABSTRACT - This section reflects the System analysis phase of the MOD project. Having established output requirements, and having characterized input, the hardware and software necessary to operate the system are specified. The General considerations portion of the section is directed to those who have little background knowledge of computer science/technology, and attempts to give a basic orientation.

"He who desires the ends desires the means."

Ramon Y. Cajal

6.0 GENERAL CONSIDERATIONS

A broadly based automated system designed to study the geographic distribution of infectious diseases will succeed, as previously stated, only if computer techniques can be effectively applied to it. The basic aspects of computer systems which are to be presented here will (hopefully) give insight into the applicability of these systems to the MOD program -- and the problems involved in making the application.

The electronic computer is one of the most powerful tools man has ever devised. It is being applied to evaluation or control of more and more areas of man's environment -- economics, science and technology, industry, and education -- to accomplish tasks which were formerly considered beyond the scope of human ability, or which required an ever increasing staff of people to accomplish. In addition to these many new tasks, all of us hope that the computer will provide a means to free us from mundane, repetitive tasks in the performance of many of the old tasks.

In the final analysis, however, the computer is recognized as just a tool -- a tool to be manipulated not by man's hands but by his mind. Engineers have built into computers the capability to perform, but it is the programers who actually cause the computers to perform. This combination of engineering and programing talents is integrated within the computer system as a latent talent. It is the user who actually supplies the motivation (in a sense) by presenting the problem to be solved.

The computer contains a control unit which allows it to sequence from one operation to the next while processing a large stream of calculations. Numbers stored in the computer represent either data or instructions and it is the programer's task to cause the computer to act on the numbers in the correct manner. The fact that both arithmetic and logical operations are possible has permitted computers to be used in a wide variety of applications. The single characteristic which has contributed most to their popularity, of course, is the rapidity with which they perform complex as well as simple operations.

Perhaps one of the most important contributions of computer technology is that it has forced man to state problems in entirely logical terms so that the computer can solve them. The opposite side of this coin is that any problem that can be stated logically or expressed in terms of mathematical equations can (in principle) be solved by a computer.

Much of the information presented in Section 6.0 is elementary and aimed at bio-medical personnel who have not had occasion for even an elementary consideration of computer technology. Those who are computer oriented are advised to turn to Section 6.1.

The computer performs storage and retrieval functions in much the same manner as a human being or a calculating machine. The computer consists of large blocks of equipment ("hardware") containing many transistors, tubes, and other basic electronic components. Most computers are organized to handle five basic functions: (1) input, (2) storage, (3) control, (4) processing, and (5) output. Before solving a problem, the pertinent facts and data must be <u>input</u> (by means of electro-mechanical devices such as card or paper tape readers, keyboards, etc.) and <u>stored</u> (on tapes, disks, drums, cores, etc.) much as a human being gathers facts and stores them in his brain. Once stored, a control section selects data, one item at a time, and <u>processes</u> it in its arithmetic element. The <u>control</u> function is simply the means of following instructions precisely as programed. The computer must be instructed (programed) every step of the way. Results are useful only after they are <u>output</u> (displayed by a printer, plotter, cathode-ray tube, etc.) or re-stored (back in memory, punched on cards, or communicated to remote devices) for later use.

We have mentioned that the computer must be programed to acquire an ability to solve problems. This is because it is impractical to build a computer capable of interpreting the wide variety of instructions that a human being could understand. And this is the reason that computing procedures are broken down to a relatively few different types of instructions. Hence,

the general plan of action written out (or encoded) as a specific set of operational instructions may be long and involved, even for an apparently simple problem. (Programs are commonly referred to as "software" to differentiate them from the computer "hardware".) We have also mentioned that the computer operates on numbers, both as data and as instructions, however, the computer does not use the familiar decimal system (with ten digits), but, as a rule, uses a binary system in which there are only two bits: 0 and 1. (The next number larger than 1 is, therefore, 10.) These characteristics, which make the computer so extremely flexible and versatile, also make it necessary for all actions to be defined in great (precise) detail. It is because of this that a major part of the human work involved in solving a problem on a computer is in preparing the program. Frequently, many man-hours are required to prepare for a few minutes of actual computer operation.

Because programing directly in machine language is tedious as well as time consuming, computer manufacturers supply (with their machines) programs that interpret so called interim languages (which are much easier for the programer to use), programs that convert this interim language program into a machine language equivalent program. A simple example is the assembler which translates sequences of characters into other sequences of characters and, in so doing, puts together, i.e., assembles, a program. For example, the characters "ADD", representing the function of addition in the programer's language, can be changed into the binary configuration (which might be "111011") that actually causes the computer to perform addition. In a sense, the assembler is acting as an interpreter. But this still leaves the programer many tedious operations. To provide further relief, more sophisticated methods have been developed. These methods usually involve a higher level interim language, approaching more closely natual English. Such language is then interpreted by a special program that translates it into a machine language, going beyond the one to one stage since several commands may have to be specified for each input

language expression. A program of this sort, which produces other programs, is called a compiler.

Since computers are designed for a variety of uses, ordinarily, no two computer manufacturers (often no computers of different type produced by a single manufacturer) produce machines which use the same internal codes to represent commands, thus a program written in an assembly language for one computer will not operate directly on another computer. This is another reason that higher level languages have been implemented on a variety of computers. (As we have implied, they also permit faster and more efficient programming.) There are several higher level languages, since problems fall into reasonably well defined categories, each of which requires a different method for solution. For example, languages for mathematical applications have been developed which are quite different from languages for business applications.

Today, techniques for producing computers are far ahead of techniques for using them. The technique of step-by-step coding of programs is wasteful of time, money, and personnel. Soon, perhaps, the computer itself can be directed to do much of the work of coding, i.e., automatic coding of programs will be possible. But until that time comes, we must accept the fact that it is time consuming and costly to produce computer programs. Because of the intimate relationships between software (programs) and hardware, a critically important part of system design is the selection of hardware. Of course the equipment must be capable of meeting task-requirements, but it should also be of such design as to minimize the amount and complexity of software that will be necessary to operate the system

6.1 HARDWARE REQUIREMENTS

The basic considerations in determining computer hardware requirements are:

continued next page

(1) The overall amount of data to be stored and the required frequency of access.

(2) The amount of data which must be considered at the same time.

(3) The method by which the data must be processed.

(4) The form in which the data must be input.

(5) The form in which the output is desired.

(6) The cost (including that of the required software).

6.1.1 OUTPUT DEVICES

Output devices are a primary consideration since the MOD system is centered around output, and we shall consider these first.

The input-output equipment of a computer is sometimes referred to as peripheral. If operated and controlled by the computer itself, it is on-line; if operated independently of the computer, it is off-line. In relatively slow computer systems the peripheral equipment is frequently on-line, but, to avoid holding up an expensive fast computer for time-consuming input-ouput operations, off-line techniques are often used. Any of the three output devices described below can be operated either on- or off-line.

(1) Line-Printer: In essence this is a very large rapid typewriter roll that prints an entire line of at least 100 characters, virtually at one time. In any case, the entire line is composed within the controller prior to printing. While a line-printer is relatively fast (up to 1000 lines per minute), it has a limited character set (usually fewer than 64), and prints letters only in upper case. It can, however, produce a rough map conveniently, rapidly, and inexpensively, since almost all computer installations have access to a line-printer. In addition, the line-printer provides the most advantageous way of generating hard-copy reports directly and rapidly.

(2) Plotter: This is a marker (pen of one color ink, a group of pens of various color inks, or a scribing point) that is mounted on a self-propelled

movable stand which draws patterns on drafting material (blank or gridded paper, or drafting plastic, or a pre-printed base map). Plotters may be operated on-line, but to conserve computer time, are more commonly operated off-line (by a magnetic tape produced on a computer, then taken off the computer and put onto the independent control unit of the plotter, which is separate from the computer). Plotters tend to be relatively slow; they may take up to several hours to draw a moderately complex map, however, they provide the highest resolution maps. A plotter commonly is limited to drawing straight lines from one point to another, nevertheless, it is a very flexible instrument since any curved line can be composed of multiple short straight line segments. Furthermore, lettering is easily performed since characters may also be composed of short straight line segments. Plot size is the primary limitation of plotters. One type, called a flatbed plotter, employs a flat drawing board with sharp limits of both width and length. Another type, called a drum plotter, utilizes a cylindrical drum around which the drafting material is wound; width is limited but not length. A plotter in the low- to medium-price range should be capable of providing finished maps, including legends, up to 30" in width.

(3) Cathode-Ray Tube (CRT) Display: This is a vacuum tube, similar to a television screen, in which a beam of electrons can be focused to a small point on a luminescent screen and varied in both position and intensity to form a pattern. A CRT operates on principles entirely similar to those which have been described for a plotter, but an electron beam is substituted for a marker pen and electronic control is substituted for electro-mechanical control. Whereas the plotter produces hard-copy output directly, the CRT screen must be photographed to obtain a lasting image. The CRT is very much faster than the plotter, but offers considerably less resolution (at its present stage of development). Furthermore, the cost of a CRT capable of meeting MOD output requirements would be prohibitive. Present MOD requirements can be satisfactorily met by using a line-printer to provide rapid map output -- for an over-view evaluation -- and by using a plotter to produce high resolution maps when these are required.

6.1.2 INPUT DEVICES

Input devices are of secondary importance, but important, nevertheless, since there is, potentially, a very large volume of data to be used in the MOD system, data which will cover all pertinent disease/environmental situations. Input devices (as are output devices) are often off-line to the central computer system; a medium such as magnetic tape serves as an intermediary. The critical problem of input lies with the conversion of raw data into a form that is acceptable to computer input devices. Input of queries into the MOD system is a closely related problem and sufficiently similar that a solution of the one should also satisfy the other. Potentially useful input devices include:

(1) Optical Character Recognition (OCR) Device: This is a device by which text can be read directly from documents and automatically translated into machine language for direct input to the computer. OCR would be of great advantage if entire reports were to be read, but in the MOD system, scientific papers are only the background material for preparing data points. Extracted data (recovered throughout the entire paper) must be converted to suitable computer input. Transcription of the extracted data could be accomplished by typing the data with a special font typewriter for optical reading or by keypunching the data onto punched cards. Either of these methods seem preferable to the use of OCR devices -- at the present time. (The high cost of OCR devices and the present rather limited state of the art were also considered in arriving at the above conclusion.)

(2) Punched Paper Tape Reader: This is a device for converting information on paper tape, punched or otherwise marked, and transfering it, one character at a time, to the computer. Paper tape is a relatively old and well standardized medium that was developed to permit more efficient use of the telegraph line. Tapes could be produced by a typist, at the typist's own rate, and the information contained on them transmitted to and from punched paper devices at the maximum rate of the transmitting and

receiving equipment. Holes punched in a moving strip of paper represent alphanumeric characters. Multiple channels on the paper tape (usually eight) are read simultaneously, permitting an entire character to be read at one time. Punched paper tape has been adapted to computers and is used extensively, particularly in less expensive systems. The paper tape reader is the least expensive type of input device.

A major disadvantage of paper tape is its inflexibility. Once it is punched, alterations, including insertions, is not possible. Correction is possible as the tape is prepared initially, but this does not suffice for the MOD system since there is need to modify data as it is input and processed. Furthermore, the paper tape does not offer the "unit-record" capability provided by other input media, and this would be a serious limitation to the MOD system.

(3) Punched Card Reader: This is a device for sensing holes punched in cards and translating that information into a form acceptable to the computer. For our purposes we can limit the discussion of punched cards to the widely used 80-column IBM punched card with its Hollerith coding representation. The punched card well reflects the unit-record concept. Each card contains 80 columns (characters) of information which can be altered without affecting the other cards in the complete record. The physical characteristics of the card facilitate sorting, collating, and other data-handling operations. Because punched cards have been in use for a long time (since the late 1800's), much auxiliary non-computer machinery has been developed to handle them, e.g., keypunch machines, designed to record data on a card in the form of punched holes, in response to an operator who strikes the appropriate keys of a typewriter-like keyboard.

(4) Digitizer: This is an analog-to-digital converter device in which the operator moves a pointer/sensor along a curve or to a point on the drawing board and presses a button, whereupon the digitizer machine reads the (X,Y) coordinates of the sensor, transmitting these directly into the

computer or onto another medium (e.g., punched cards or magnetic tape). The Army Map Service and the Bureau of the Census both utilize digitizers in their work. In the MOD system, digitizers could be quite useful in translating data from existing maps into computer form for use in subsequent processing.

6.1.3 STORAGE DEVICES

All computers have a rapid access storage device, usually randomly accessible. It is in this primary, i.e., main, storage device that program instructions and data are stored and from which instructions are retrieved by the control unit and executed. Main storage is of interest only in that there is a mimimum requirement (to be discussed later) for the MOD system.

In addition to the computer's main storage, auxiliary storage devices must be provided in which MOD data of all types will be filed. Auxiliary storage has a much greater capacity than main storage, however, the information is less rapidly accessible. Three of the various types of auxiliary storage devices are described below.

(1) Magnetic Drum: is a rapidly rotating cylinder whose outer surface is coated with magnetic material. It provides moderately rapid random-access storage. It is expensive.

(2) Magnetic Disk: is a stack of rapidly rotating flat disks, having their flat surfaces coated with magnetic material. It provides a moderately rapid, random-access storage It is moderately expensive.

(3) Magnetic Tape: is a steel or plastic tape coated with magnetic material and wound on a reel. It provides slow sequential-acess storage (however, data can be arranged initially -- by card-sorters -- or later -- by the computer itself -- so that long, slow random searches will seldom be necessary. It is inexpensive.

6.1.4 CENTRAL PROCESSING UNITS (CPU)

The central processor (CPU) of the computer system normally consists of the main storage, arithmetic unit, control unit, and special register groups. It is the principal unit of the computer; it controls the processing routines, performs the arithmetic functions, and maintains a quickly accessible memory. The design characteristics of a computer are most noticeably reflected by the CPU. CPU considerations of the MOD system which affect the selection of a computer deal with the processing commands (steps) designed into the computer, the processing speed, and the input-output interfacing. Some computers are much more appropriate than others for solution of particular problems, due to the amount of main memory and the internal processing speed (influenced by the complexity of the machine commands). The amount of input-output performed can greatly influence operation unless techniques are provided to avoid interruption of processing while input-output operations are performed.

6.1.5 AVAILABLE SERVICES

There are three ways in which the MOD project staff could satisfy its computer requirements

(1) Purchase its own computer.

(2) Rent or lease a computer to be installed on the AFIP premises.

(3) Rent time on a computer installed on another organization's premises (either a nearby government agency, or a cooperating university such as the University of Illinois). The Computer Sharing Exchange (part of the General Services Administration) maintains a record of all government computers in the Washington, D.C. area on which time would be available at a nominal cost. Computer manufacturers also represent a potential source of computer time in the Washington, D. C. area.

6.2 SOFTWARE REQUIREMENTS

There are several types of languages available for use in programming the MOD system. The choice is somewhat dependent upon the computer hardware selected because many languages have been implemented for only a limited number of computers. In general, low level, or assembly languages will not be used, as the time and effort involved in programming in these languages is greater than that required when a more universal compiler language is used. The most widely used and generally applicable of the higher level languages are described below.

6.2.1 AVAILABLE LANGUAGES

Each of the available languages was developed by an individual or a group to satisfy a very general need in a particular type of application. It is for this reason that we must consider several of the higher level languages. These languages are procedure-oriented and machine-independent. None of these languages can be executed directly by present computers without first being "processed" into machine language, but this is their advantage. This design allows them to be implemented for a variety of computers with basically no changes in the language itself.

(1) ALGOL: was developed in Europe to be an internationally accepted procedure for designing mathematical, engineering, and scientific problems. Compatible standardization and understanding of problems and procedures to be used with or without computers were the primary objectives in developing ALGOL. The language provides precise instructional statements and ways of expressing problem-solving order and procedure. The ALGOL language (or abridged subsets of it) is currently available for a limited number of computers.

(2) COBOL: was developed by a consortium of computer manufacturers and users (including groups within the Federal Government). It grew out of

a desire for a language that would be a "shorthand" for computer instruction, yet derived from English. It resembles English, so the programmer can work with it easily without having to learn many special symbols and codes, and special rules for using them. Instructions written in COBOL are sentences that are meaningful even to the casual reader. The COBOL language is capable of describing business problems of many kinds and can easily specify the basic steps required to solve them.

(3) FORTRAN: was developed by IBM and is presently the most widely used language for scientific problems and programs. It was developed in the United States in parallel with ALGOL (in Europe). Both languages are attempts to provide a programming language similar to everyday mathematical notation so that engineers and scientists can avoid the repetition and drudgery of machine programming -- and both succeed. Newer versions of FORTRAN include features formerly available in ALGOL alone. The grammar, symbols, rules, and syntax used are, for the most part, easily learned since they follow conventional mathematical and English-language usage, but the instructions must be explicit.

6.2.2 AVAILABLE SERVICES

There are two ways in which the MOD project could obtain the services necessary to implement the required programs:

(1) Hire its own programming staff.

(2) Contract the programming tasks to a professional data-processing organization.

6.3 CONCLUSIONS

The MOD system's output can and usually will be in the form of maps drawn off-line by an ink-on-paper plotter. For making interim maps, a high-speed printer or a CRT-microfilm plotter could be used, however, because of the limited selection of characters available on the former,

and the limited precision (on a single plot) of both, maps produced by either of these devices are likely to be of significantly lower quality than those produced by an ink-on-paper plotter. Other output media are either inapplicable, too inflexible, too slow, or too expensive for our purposes.

Out of the vast array of possible input devices, it seems most practical for the MOD system to adopt the widely-used punched cards and magnetic tape for input, though possibly, digitizers may prove useful to input data which is already in map form. The nearly universal use of punched cards and magnetic tape has resulted in a substantial body of equipment and experience which will be of great value in handling these two input media. The other media are not applicable in this project because they are not sufficiently flexible, or too slow, or not sufficiently developed to be practical at this time.

At present, we believe that sequential-access (magnetic tape) storage will be adequate for the MOD system initially (in addition to the direct-access main storage element of the computer itself). Later, however, it may prove desirable to add random-access (preferably magnetic disk) storage to the system. The size of main memory is a much more important factor in mapping requirements than it is in data storage and retrieval requirements for the following reason. In order to construct contoured or shaded maps a grid must be employed. A very general trend map can be prepared by utilizing a 10 x 10 grid (100 points), alternatively, a fairly detailed map can be prepared by utilizing a 100 x 100 grid (10,000 points). Since each point consists of three values (X,Y, and Z), and one computer word is required for each value, main memory must contain at least 30,000 words to produce a "more detailed" map. This requirement, plus the main memory requirement for storage of the computer program itself, brings the total main memory requirement to approximately 50,000 words.

The tasks to be performed by the CPU in the MOD system are primarily of a logical rather than a mathematical type. While speed is not a prime factor in the information storage and retrieval tasks, it is a factor in the processing of the hundreds of thousands of data points which are required in mapping. This means that the computer selected must represent a compromise between one designed for information storage and retrieval tasks (usually a small, slow machine) and one designed for general scientific tasks (usually larger and faster). Alternatively, two types of computers could be selected -- one for performing the information storage and retrieval tasks, the other for performing the mapping tasks. (During the early efforts to implement the MOD system, this alternative method was used to good advantage.)

The computer time used in the design and implementation of the MOD project was, for the most part, rented on available computers or obtained (gratis) from the Computer Sharing Exchange. Time was rented from the Control Data Corporation in order to use their contour mapping program system. (We used the CDC 3600 and 160-A computers and the CalComp 564 plotter.) We were permitted to use several grovernment computers on a non-interference basis, including the IBM 7090 of the Strategy and Tactics Analysis Group (STAG), the IBM 7094 at the National Aeronautics and Space Administration (NASA), the IBM 7090 at the Naval Command Systems Support Activity (NAVCOSSAC), and the CDC 3100 at the Naval Oceanographic Office (NAVOCEANO) -- and we are most grateful for this opportunity. Computer programs were provided by the Kansas Geological Survey and NAVOCEANO. Some maps were produced for us by the University of Michigan on their IBM 7090. We also wrote some of our own programs, and these were used at NASA, and NAVCOSSAC, and at the AFIP computer center (which contains an IBM 360/30).

Software requirements can be met by any of the systems described, but there are other factors to be considered -- see 6.2.1. For example, ALGOL is perhaps the least available language (in the United States, but

not in Europe). COBOL is an easy language in which to work, but does not have the scientific capability of FORTRAN. On the other hand, FORTRAN has a rather limited data processing capability, especially the versions implemented by IBM. CDC FORTRAN has perhaps the best overall capability for programming the MOD system, but is only available for CDC computers. Most programs which we borrowed or purchased for map construction were written in FORTRAN. (The exception was at the University of Michigan where the MAD language is used).

We have utilized both methods of obtaining programming services: by hiring H.M.Kline, a computer analyst-programmer, and by contracting with Planning Research Corporation for programming (as well as for system analysis and design).

Conclusions from system analysis indicates that the MOD computer system should be capable of performing the following functions:

(1) Input and edit data.

(2) Generate data files employing the input data.

(3) Input and edit queries.

(4) Retrieve disease/environmental information from the data files based on the query set.

(5) Perform high-speed sorts.

(6) Calculate, using mathematical functions.

(7) Generate commands for an automatic data-plotting device.

(8) Generate auxiliary hard-copy (printed reports).

(9) Display contents of any portion of the data files.

These requirements indicate that a medium- to large-scale computer is required for the final system. Design studies have shown ways by which an interim implementation can be carried out on a small-scale computer (such as the IBM 360/30 at AFIP), requiring only map generation to

be performed on a larger-scale system (such as the IBM 7090), because existing programs require such a computer configuration. It would be possible to convert all programs developed for such an interim system (if they were written in COBOL or FORTRAN) into a common system for use on a large-scale computer. An off-line plotter can be used to produce all maps -- and such plotters are readily available in the Washington area.

7 Data processing

ABSTRACT - This section reflects the System design phase of the MOD system. It considers in detail the various subsystems:

Storage subsystem
Retrieval subsystem
Synthesis subsystem
Output subsystem

discussing their structure and their function. Flow diagrams are presented.

". . . a discovery is nothing more than the union of two or more truths to a useful end."

Ramon Y. Cajal

7.0 GENERAL CONSIDERATIONS

Data Processing is commonly defined as the "rearrangement and refinement of raw data into a form suitable for further use" or "any procedure for receiving information and producing a specific result" (Sippl, 1966, p. 88). Automatic data processing is "data processing performed by a system of electronic or electrical machines so interconnected and interacting as to reduce to a minimum the need for human assistance or intervention" (Data Nation, 1966, p. 40). Thus Data Processing includes all operations necessary to produce the desired output (results) from the available input (data) utilizing the selected computer hardware. These operations include whatever manual interfacing with the computer is required to input data, corrections, and requests into the system as well as those tasks which can be performed by the machines under the direction of suitable internally stored programs.

The information storage and retrieval (IS&R) portions of the MOD system had been completely designed at the time when work on the project was terminated. Because of this, the techniques for building, storing, maintaining, and retrieving the MOD data could be specified, despite some unresolved input and output problems, since these latter aspects are independent of explicit input and output considerations once the essential elements of a computerized system have been determined.

In the MOD system the essential element is the formulation of data points, an important part of which consists of a LOF/MOF structure. For the purpose of MOD data processing, LOC, VAL, and NAR of a data point can all be treated in essentially the same manner as MOF's. These data points must be stored and retrieved in the MOD system regardless of the manner in which extrinsic problems may later be resolved. Moreover, all of the possible mapping problems which might be encountered in using the MOD system cannot be anticipated until an attempt is made to produce maps by using actual MOD data. In particular, the acceptability of existing computer mapping methods and programs cannot be ascertained a priori. The LOF/MOF structure of the

input (and implicit in retrieval re, ests) is open-ended, not only to afford maximum flexibility to the system, but also because not all pertinent descriptive elements can be predicted at the outset.

In the MOD system it is anticipated that additions of new MOF's and redefinitions of existing MOF's and LOF's are quite likely. Precise requirements for such restructuring will become apparent only when actual data are used as input into the system and real retrieval attempts are conducted. Therefore, in a modular approach to the design and implementation of the entire MOD system, the IS&R subsystems represent a logical building block and test tool for the remaining facets of the system.

Because of this, the following sections provide detailed design specifications for the Storage and Retrieval Subsystems while the Synthesis and Output Subsystems are treated in a more general fashion. The descriptions of the first two subsystems contain specifications for their immediate program design and implementation. These subsystems have been designed so that subsequent modification to them should be unnecessary. However, the rationale for the techniques and methods utilized in those subsystems are given so that if changes seem desirable it will be easier to evaluate their feasibility -- and complexity.

The designed system is applicable to either a magnetic tape or disk computer configuration. But special considerations were given to the additional processing which would be required in a tape system since, during the design phase, it appeared that system implementation would be with non-random access files.

The formats for the various input cards are provided not only to complete the design specifications of the Storage and Retrieval Subsystems, but also in order that the data can, if desired, be collected and transcribed onto cards simultaneously with future program development.

Two functions of the Dictionary File are so intimately involved in the synthesizing operations necessary to produce reports and maps from MOD data that discussion of these functions is deferred until the Synthesis Subsystem is described, although these two functions - gazetteer and grid - could also be considered logically under the Storage Subsystem.

One of the most important programs to be written eventually is a control program which will coordinate the operations of the various MOD subsystems. This control program will read all the control information and determine the proper subsystems to be called in at the appropriate times. It will minimize possible procedural errors (and the necessity for computer operator intervention) and maximize efficiency of total system operation. Design of the control program will be based upon the finished subsystems, for which reason this program will be the last one to be designed and implemented.

The various components of the MOD system design are graphically summarized in the overall functional chart of the system (Fig. 7.1).

Concrete examples are provided wherever possible to demonstrate and amplify the abstract discussion. These examples are accurate and realistic for illustrative purposes, but they are not necessarily exhaustively complete, lest they become unmanageable.

7.1 STORAGE SUBSYSTEM

7.1.1 DATA INPUT CARDS

The data contained on the data extraction forms are entered into the MOD system by means of punched cards. An attempt has been made to allow the data to be keypunched as it appears on this extraction form, with as few additional instructions to the keypunch operators as possible. The preprinted MOF designation, including its surrounding parentheses, is punched for each MOF utilized. (These parentheses makes the cards more readable

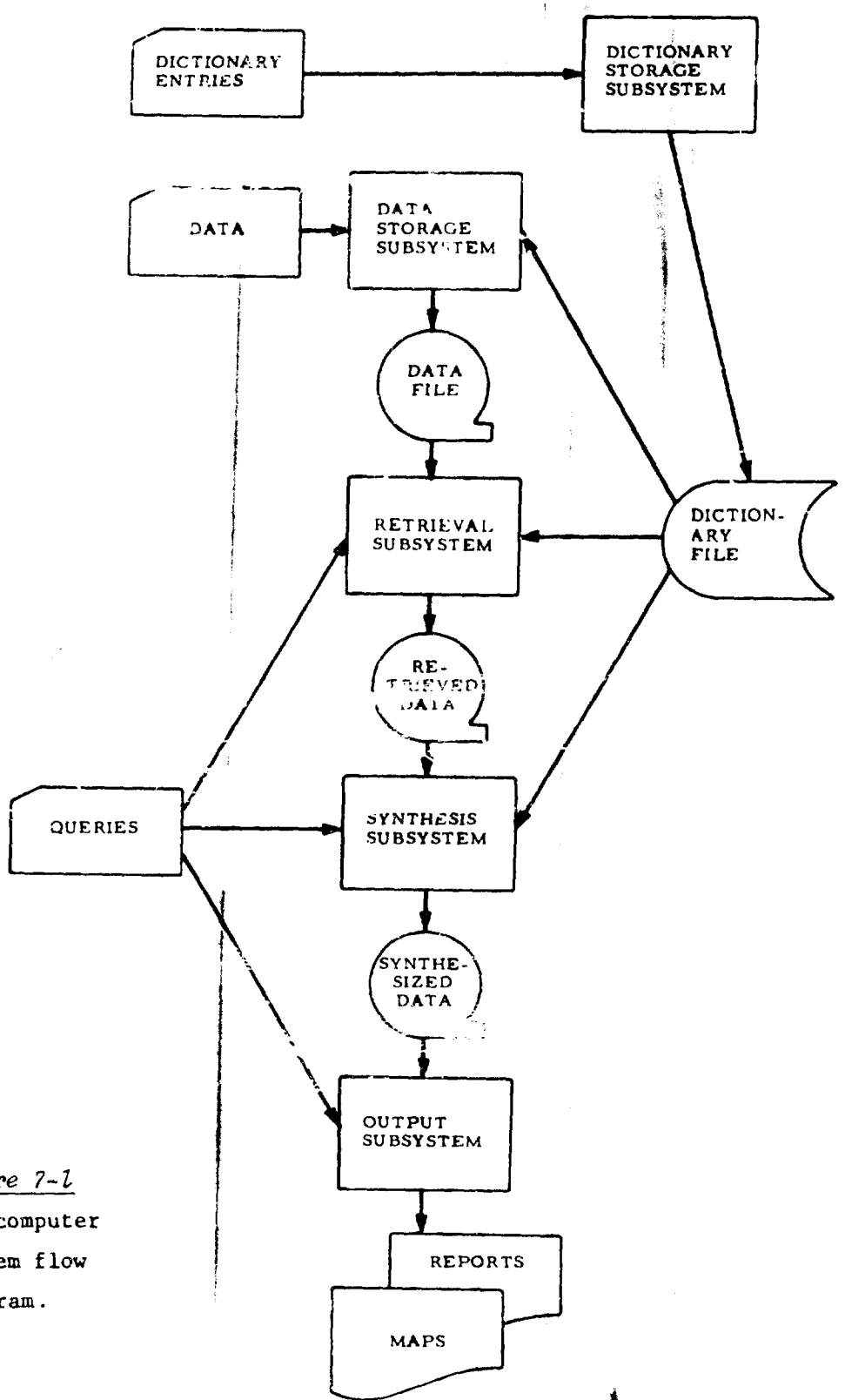


Figure 7-1
MOD computer system flow diagram.

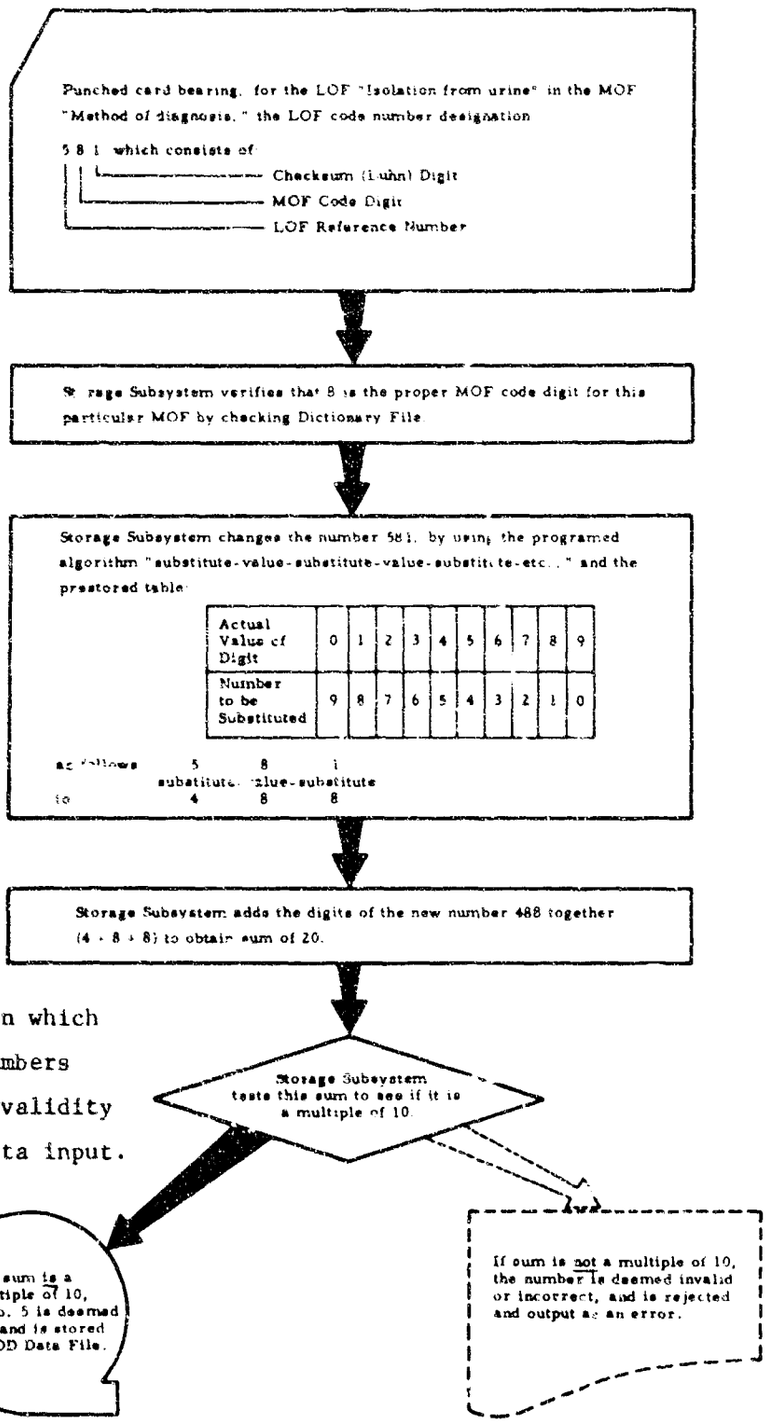


Figure 7-2 Manner in which preprinted LOF code numbers function in automated validity checking during MOD data input.

f manual verification.) Each LOF associated with this MOF is then punched as it appears on the data form. A listing of some MOD data input cards has already been given (Fig. 5.8).

LOF's may appear on the extraction form as preprinted code numbers, written numeric quantities, written code numbers, and as written textual spellings. For those MOF's containing LOF's which are numeric values (i.e., "quantitative" LOF's), the numerals are always handwritten on the form. However, for MOF's which contain an open-ended set of predefined LOF's (alphabetic or "qualitative" LOF's), the LOF's may appear either as the preprinted reference number appearing on the data form, as an additional reference number supplied by the data analyst, or as the textual spelling of the LOF written by the data extractor. In order to minimize the possibility of otherwise undetectable keypunch errors, the preprinted LOF numbers actually consist of the LOF reference number, a MOF code digit, and a checksum (Luhn) digit; this is shown in Fig. 7.2. In those MOF's which may be specified by several LOF's, (which, hereafter, are called multi-LOF MOF's), the LOF indications on the data form will contain preprinted commas which will also be keypunched onto the input cards to separate these LOF's. For MOF's which must be specified by a single LOF (which, hereafter, are termed single-LOF MOF's), there will be no preprinted commas. Vague or questionable data may be marked with a "?" on the data extraction form and such question marks will be keypunched immediately after the pertinent LOF's.

These data input cards, which are to be used for both initial data entry and subsequent data maintenance, have the following format:

- see next page -

Identification Field	card columns 1-13	
Data Point Number		
Year of Extraction		1-2
Month of Extraction		3-4
Day of Extraction		5-6
Extractor's Identification (EID)		7-9
Data Point Number (that day)		10-12
Card Type (if required)		13-13
Data Field	14-80	

The data field contains the MOF and LOF data in free form, hence it has no predetermined subfields. Blanks not embedded within a textual spelling of a LOF are optional between entries. Thus any card could contain only one MOF and one LOF, or as many as eight MOF's if each MOF contained only one preprinted LOF code designation.

Although one card can be punched without special instructions to the keypunch operator, special instructions are necessary to handle continuation cards for a data point. These instructions (rules) also reduce the amount of preliminary processing required:

(1) Each card must contain the data point number and card type in columns 1-13.

(2) A LOF must be entirely contained on a single card (whether numeric, code reference number, or textual spelling).

(3) If a new LOF of a previously designated MOF is to be placed on a different input card, the MOF designation must be repeated.

These requirements limit the length of a LOF to 62 characters (67 characters in the data field minus 5 characters for the MOF designation). As shall become evident, this size limitation proves convenient for the Dictionary File. Note that, for input purposes, NAR (narrative) of a data point can be treated as another MOF -- but the "MOF" for NAR has no size

restrictions. It is suggested that each narrative card contain a continuation card number in the data field immediately preceding the MOF designation (i.e., starting in card column 14).

File maintenance cards have the same format as the original data cards. The type of maintenance to be performed is indicated in the card-type field, card column 13, by the following codes:

D - Delete all MOF's indicated on the card, or if no MOF is indicated, delete the entire data point record.

R - Replace the LOF's of the indicated MOF with the listed LOF's. This operation is a strict replacement; if only one of a series of LOF's for a MOF is to be changed, all of the immutable LOF's must also be indicated on the replacement card if these changes are to be effected in one computer pass.

A - Add the designated LOF to the indicated MOF. If a LOF already exists for this MOF, the new LOF will be added to the existing LOF('s) if the MOF may have several LOF's. A blank in the card-type field is utilized on initial entries. Each card, obviously, can contain only one maintenance code although several MOF's may be specified on the one card.

From this it is seen that all data cards contain a MOF designation in card columns 14-18, with the possible exception of narrative (NAR) cards. All types of cards may contain as many or as few MOF's and LOF's as are consistent with the rules given for continuation cards.

Normally, file maintenance and creation of a given data point will be performed at separate points in time. However, if several types of entries are processed at the same time for the same data point, they will be considered in the following order of precedence:

(1) Delete; (2) Replace; (3) Add; (4) Initial Entry

7.1.2 DATA FILE

All of the MOD input data which has been accepted by the MOD system are stored in the Data File. This file consists of one logical record for each data point. Each record contains the data point number and all MOF's and LOF's which pertain to that data point. Even in coded form these data are quite variable since there are common and optional MOF's, since some MOF's may contain several LOF's, and since the data may include an unspecified amount of narration (NAR). For these reasons it is impractical to utilize fixed length records for a data point.

All LOF's (except material appearing under the narrative, NAR) will be represented in the Data File by numbers. A "?" will be appended to any LOF entry for which the input data was so marked. The actual values of quantitative LOF's will be used, however, as shall be seen in the next section, qualitative LOF's will be represented by a code number the size of which depends upon the number of levels in its generic tree structure.

Each LOF must be associated with its appropriate MOF. This could be accomplished in several ways, e.g., each LOF or group of LOF's could be immediately preceded in the record by its MOF designation. There is a serious disadvantage to this solution because the entire record would have to be searched to locate any given MOF. A more desirable method is to create an index within each record which would establish the relative location (within the record) of the first LOF for each MOF. This would require that the length assigned to each LOF be provided since length varies from MOF to MOF. The index itself could be either fixed or of variable length since not all MOF's are present in each data point (but sufficient locations could be set aside to provide for all presently defined MOF's in the index). A fixed length index would reduce somewhat searching requirements, however, many programs of the MOD system would have to be updated in order to process a new index structure when new MOF's were added to the dictionary. Because of this complication we have chosen to make the index

variable and to consist of the MOF designation, relative starting location, and length of LOF code for each MOF.

In general, the MOF designations will be arranged in alphabetical order in the index, and the LOF representations will also be sequenced in this MOF order. The narrative (NAR) should appear last in the logical record and, perhaps, even in a separate physical record. It would be desirable to place certain fixed-length single-LOF essential MOF's (e.g., the geographical location (LOC) and the value (VAL) for the data point), in a fixed location within each data point record for facility in sorting and other manipulations. If this were done it would not be necessary to include these MOF's in the index.

The format of each data point logical record can then be described as:

(1) Data Point Number -- fixed location, format, and length.

(2) Predetermined essential MOF's, LOC, and VAL -- fixed location, format, and length.

(3) Record index of other MOF's -- fixed starting location and format, but variable length.

(4) LOF's -- variable starting location, format, and length.

(5) Narrative (NAR) -- variable starting location, format, and length.

Design of the MOD system has been based largely upon two disease "models" for reasons discussed in Sections 1 and 2. However, virtually an unlimited number of diseases could be processed by the system. This would require only design of new data extraction forms and the selection and definition of new MOF's and additional LOF's (even for previously existing MOF's). There would seem to be no requirement for maintaining a different dictionary for each disease although it might prove desirable to place different disease data in separate data files.

As in the case of diseases, an almost unlimited variety of environmental data could be processed by the MOD system -- with appropriate new data extraction forms, MOF's, and LOF's. The present system is capable of drawing environmental maps, but will seldom be used to do that. Instead, the scale and projection of MOD-produced disease maps will be adjusted to correspond with those of existing environmental maps so that the one may be readily compared with the other. Environmental factors extracted along with disease data will be considered only with the data point for which they are included as MOF's hence their output capacity (under these conditions) will be restricted to retrieval functions. However, a data file of environmental factors could be built from either or both the input disease data and that derived from separate environmental data extraction forms. Environmental data points generated by the former means would contain the associated medical data point number; those by the latter would not be associated directly with particular disease data points. In addition, it would be possible to produce single environmental factor files by techniques which would digitize existing maps.

From these considerations it is evident that environmental maps could be produced from the MOD data files in which each data point was obtained from environmental data (if present) or disease data. In a retrieval which included both environmental and disease conditions, the user could specify that only those factors explicitly associated with the disease should be considered. On the other hand he could broaden his retrieval to include corresponding factors from the environmental files or even those from other disease files.

As an illustration of the foregoing, if the Data File contained internal codes equivalent to the following data:

```
670828JDS004 (LOC) POPE CO., ILLINOIS
             (VAL) 2
             (TIB) 621225
             (TIE) 630228
             (SEA) SUMMER
             (TOD) MORNING
             (SSZ) 10
             (LSZ) 17
             (SDA) L. GRIPPO, L. BALLUM
             (MDG) ISOLATION FROM TISSUE
             (NAR) TRANSMSN POSS PREDATION ON FERAL HOUSE-MOUSE
                   TITERS: LBALL=1:100
670828JDS005 (LOC) JOHNSON CO., ILLINOIS
             (VAL) 12
             (SDA) LEPTOSPIRA
             ----------> - then the following cards:

670828JDS004R(TOD) DAWN, DUSK (SSZ) 9 (SEA) WINTER
670828JDS004R2(NAR) TITERS:  LHYOS=1:1000; LBALL=1:100
670828JDS004A(SDA) L. HYOS
670828JDS005D
670829JDS001 (LOC) MASSAC CO., ILLINOIS
             (VAL) 5
             (TIB) 63
             (TIE) 63
             (TOD) AFTERNOON, DUSK
             (SSZ) 37
             (LSZ) 37
             (SDA) L. CANICOLA
             ----------> - would cause the Data File to
                           contain data equivalent to:

670828JDS004 (LOC) POPE CO., ILLINOIS
             (VAL) 2
             (TIB) 62-12-25
             (TIE) 63-02-28
             (SEA) WINTER
             (TOD) DAWN, DUSK
             (SSZ) 9
             (LSZ) 17
             (SDA) L. GRIPPO, L. BALLUM, L. HYOS
             (NAR) TRANSMSN POSS PREDATION ON FERAL HOUSE-MOUSE
                   TITERS:  LHYOS=1:1000, LBALL=1:100
670829JDS001 (LOC) MASSAC CO., ILLINOIS
             (VAL) 5
             (TIB) 63
             (TOD) AFTERNOON, DUSK
             (SSZ) 37
             (LSZ) 37
             (SDA) L. CANICOLA
```

7.1.3 DICTIONARY FILE

The Dictionary File is the central element of the MOD system. It is the link which connects input data to stored data to retrieved output data. This file contains the dictionary of terms which are allowed as both data descriptors and query descriptors, thus the dictionary is also a bridge between the environment of the medical doctor and the environment of the computer.

A logical record in the Dictionary File contains descriptors for a complete MOF, i.e., both the MOF description and all its associated LOF descriptions. The MOF description consists of the complete English language description (long form) of the MOF, and the abbreviated description (short form) consisting of three letters, e.g., Specific Disease Agent: SDA. The LOF description consists of the English language description of the LOF and the LOF code number.

In order to keep the internal data consistent, but to allow freedom of synonymous expression externally, synonyms and variant spellings can be incorporated in the Dictionary File. Variant spellings will be corrected as the data is input so that all printouts will contain the preferred form of each LOF. Synonyms are keyed to the preferred form, but are carried internally with their own identifier. This permits a query on a group of synonymous terms or on the specific term requested (called synonym lockout). As a further convenience to the query requestor, terms which fall within a category are automatically provided. This is accomplished by means of an internal tree structure of terms. For example, a query on "mice" would yield data on Muridae, also on each of the members of the family: Mus musculus, Pitymys, etc.

The Dictionary File contains all of the MOF's and LOF's which have been defined for the MOD system and has been designed for either magnetic tape or disk storage. A MOF is considered as defined in the MOD system if

the Dictionary File contains record for that MOF. MOF's should be so defined that no MOF contains both quantitative and qualitative LOF's. A quantitative (or numerical) LOF is considered as defined if it has a valid numeric value. A qualitative (or alphabetic) LOF is considered as defined if the dictionary contains a LOF entry for it.

A MOF record includes the textual spelling of the MOF name and its (short form) MOF designation. If a MOF has quantitative LOF's, the MOF record will also contain an indication of the type of edit checking to which its LOF's are to be subjected. The validity of dates and ranges can be tested in numeric LOF's. If a MOF consists of qualitative LOF's, the MOF record will also contain the MOF code digit which is associated with each LOF number on the data extraction form.

For qualitative MOF's there will be a record for each LOF thus far defined in the system. This record will contain an indication of the structural relationship of the LOF to all other LOF's with the MOF. These relationships consist of generic tree levels, synonyms, and variant spellings. Each LOF (within a MOF) is assigned a reference number. This number, which is provided by Dictionary File listings and used in updating, appears on the data extraction form along with the MOF code digit and a checksum digit. The structural relationships are indicated by the LOF code number. This code number is composed of reference numbers, one for each tree level (the reference numbers for the (main) LOF at each higher tree level plus the reference number for this LOF itself). In addition, another reference number is used for synonym designation. Thus a code number consists of a series of reference numbers, the length of the series being one greater than the total number of levels in the given MOF. Although variant spellings are separate entries, they are assigned the same reference number as the LOF for which they are a variant.

The explicit format of the Dictionary File will, of course, depend upon the computer selected and the available external storage devices. In

any event, the file should be structured to facilitate input and retrieval of the data even though this will require additional processing within the dictionary.

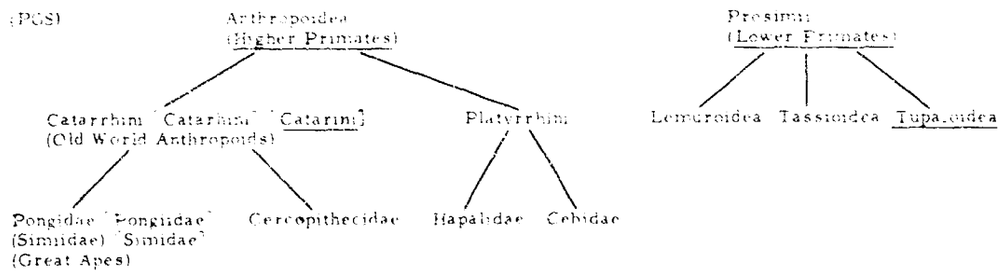
Since the input data will include both reference numbers and words, random searching within a MOF record can be eliminated if both an alphabetical order (for the words) and a numeric order (for the reference numbers) are maintained. This method will also be helpful in processing word LOF's of the retrieval requests and reference number results of the retrieval. These two sequences within the dictionary can be maintained on a disk by separating the LOF records into two sections. The alphabetical order records would contain the LOF word and reference number; the numeric order records would contain the reference number, the code number, and an index number to indicate the location of the related alphabetic order record. (There would only be one numeric order record for all variant spellings of the same word.) In this way utilization of disks is minimized. If the Dictionary File is maintained on magnetic tape, both records should contain the LOF word and code number, and in essentially the same format of LOF word, reference number, and code number. To minimize the amount of processing time required, they should be kept in different (physical) files.

The order of the MOF's themselves, within the Dictionary File, is somewhat arbitrary, however, two important factors should be considered. First, the Dictionary File maintenance cards must eventually be sequenced in the same order as the Dictionary File. Secondly, on tape, it would be most convenient to place together an entire set of LOF's common to more than one MOF. (Perhaps the order of MOF's included on the data extraction form could be arranged to facilitate this.)

The following example, which illustrates the internal structure of a MOF, will also serve as an example of MOF construction. (The synthesis section contains a somewhat similar example, but deals with geographic locations.) In this and following examples, brackets "[ ]"indicate variant

spellings and parentheses "( )" indicat synonyms.

Consider a MOF, "Primate groups involved in study (PGS)", composed of several LOF's arranged in the following tree-structure, in which the underlined LOF's are to be added to the MOF:



On magnetic tape, the MOF, (PGS), would appear as follows:

ALPHABETIC ORDER							
MOF DESC	MOF TEXTUAL SPELLING	# ENTRS			LVL	CODE	
(PGS)	PRIMATE GROUPS INVOLVED IN STUDY	13			3	2	
	LOF TEXTUAL SPELLING	LOF CODE #				REF #	
(PGS)	ANTHROPOIDEA	1	0	0	0	1	
(PGS	CATARHINI	1	2	0	0	2	
(PGS)	CATARRHINI	1	2	0	0	2	
(PGS)	CEBIDAE	1	8	10	0	10	
(PGS)	CERCOPITHECIDAE	1	2	7	0	7	
(PGS)	GREAT APES	1	2	4	6	6	
(PGS)	HAPALIDAE	1	8	9	0	9	
(PGS)	LEMUROIDEA	11	12	0	0	12	
(PGS)	OLD WORLD ANTHROPOIDS	1	2	0	3	3	
(PGS)	PLATYRRHINI	1	8	0	0	8	
(PGS)	PONGIDAE	1	2	4	0	4	
(PGS)	PONGIIDAE	1	2	4	0	4	
(PGS)	PROSIMI	11	0	0	0	11	
(PGS)	SIMIIDAE	1	2	4	5	5	
(PGS)	SIMIDAE	1	2	4	5	5	
(PGS)	TASSIOIDEA	11	13	0	0	13	

NUMERIC ORDER							
(PGS)	PRIMATE GROUPS INVOLVED IN STUDY	13			3	2 *	
(PGS)	ANTHROPOIDEA	1	0	0	0	1	
(PGS)	CATARRHINI	1	2	0	0	2	
(PGS)	CATARHINI	1	2	0	0	2	*
(PGS)	OLD WORLD ANTHROPOIDS	1	2	0	3	3	
(PGS)	PONGIDAE	1	2	4	0	4	
(PGS)	PONGIIDAE	1	2	4	0	4	*
(PGS)	SIMIIDAE	1	2	4	5	5	
(PGS)	SIMIDAE	1	2	4	5	5	*
(PGS)	GREAT APES	1	2	4	6	6	
(PGS)	CERCOPITHECIDAE	1	2	7	0	7	
(PGS)	PLATYRRHINI	1	8	0	0	8	
(PGS)	HAPALIDAE	1	8	9	0	9	
(PGS)	CEBIDAE	1	8	10	0	10	
(PGS)	PROSIMI	11	0	0	0	11	
(PGS)	LEMUROIDEA	11	12	0	0	12	
(PGS)	TASSIOIDEA	11	13	0	0	13	

* These records could be eliminated from the numeric order file.

If the Dictionary File were maintained on a disk file, the alphabetical order records for any MOF would be similar to those for a tape file with the exception that the MOF designator would only need to appear in the MOF description record. The numeric order records could contain an indication of the location of the textual description of the LOF rather than the description itself, and would appear as follows for the MOF (PGS):

1	1	0	0	0	1
3	1	2	0	0	2
9	1	2	0	3	3
11	1	2	4	0	4
15	1	2	4	5	5
6	1	2	4	6	6
5	1	2	7	0	7
10	1	8	0	0	8
7	1	8	9	0	9
4	1	8	10	0	10
13	11	0	0	0	11
8	11	12	0	0	12
16	11	13	0	0	13

After the four new (underlined) LOF's were added to the MOF, (PGS), the Dictionary File would include the following four new records (in both alphabetical and numeric order) in a tape system:

(PGS)	CATARINI	1	2	0	0	2
(PGS)	HIGHER PRIMATES	1	0	0	14	14
(PGS)	LOWER PRIMATES	11	0	0	15	15
(PGS)	TUPAIOIDEA	11	16	0	0	16

The resultant numeric order section of the MOF, (PGS), would appear as follows in a disk system, where the location references have been changed to reflect additional LOF's.

1	1	0	0	0	1
4	1	2	0	0	2
12	1	2	0	3	3
14	1	2	4	0	4
18	1	2	4	5	5
7	1	2	4	6	6
6	1	2	7	0	7
13	1	8	0	0	8
8	1	8	9	0	9
5	1	8	10	0	10
16	11	0	0	0	11
10	11	12	0	0	12
19	11	13	0	0	13
9	1	0	0	14	14
11	11	0	0	15	15
20	11	16	0	0	16

7.1.4 DICTIONARY INPUT CARDS

Dictionary File building and maintenance consists of the following operations:

- Construction (or Reconstruction) -- the entire Dictionary File, or a set consisting of several of its component/MOF's, is constructed or reconstructed to correct gross errors. In addition, entire new MOF's are incorporated into the Dictionary File by this method.
- Updating -- new LOF's are added to MOF's already existing in the Dictionary File.
- Correction -- a LOF or MOF is deleted or has its verbal description changed.

A single card format has been designed to process all of these types of file maintenance. This format provides uniformity in the coding of all dictionary cards and allows for the recreation of the entire Dictionary File, utilizing all existing cards. The same format is also used to generate the MOF description entry. The general format of these cards (in which each element is left justified) is as follows:

CARD COLUMNS	CONTENTS	USAGE
1-5	MOF three character designation enclosed by the usual parentheses.	All types*
6-6	MOF code - one digit used to verify key-punching of coded entries or special code to indicate that the MOF is processed in an exceptional manner.	MOF entries
7-72	Clear text spelling of the MOF or LOF. (Note that this spelling is limited to 66 characters.)	All types
73-74	Structure indicator (MOF entry contains total number of levels in structure).	All types
	(*MOF designation is somewhat redundant for construction of LOF entries, but is helpful in ordering the cards.)	

—— continued next page

75-80	(LOF entries only) Existing dictionary LOF reference number for referenced LOF.	Updating &
	Desired external LOF reference number if number is to be preprinted in the data extraction form.	Construction

7.1.4.1 MOF Construction (or Reconstruction) In order to create a MOF in the Dictionary File or to update a MOF to include new generic levels, the entire MOF must be constructed (or reconstructed) as a unit.

For simplicity in coding, the LOF's are sequenced by their generic level and contain a structure indicator which designates this level, or their usage as a synonym, or a variant spelling.

These indicators are as follows:

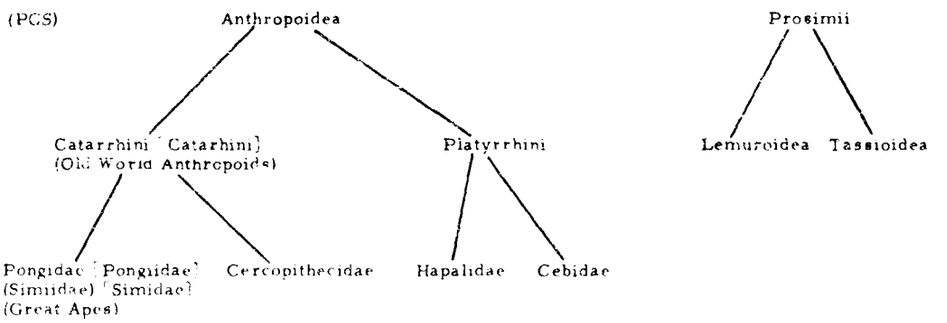
1,2,3,...	level
$	variant spelling
=	synonym

Each LOF is assigned only one indicator for brevity and ease in resequencing so that variant spellings and synonyms are considered to be at the previously indicated level. Since variant spellings pertain to a particular word whereas synonyms apply to a possible group of words, variant spellings must follow immediately their object word. Moreover, since the connotation of words cannot be considered, the sequence of a group of synonyms, including the determination of the base word (assigned the lev indicator), is somewhat optional.

If an existing MOF is restructured, both the Data and Dictionary File entries which pertain to that MOF must be recreated. Hence, when the level structure of a MOF may contain unknown lower levels, it is desirable to indicate the maximum number of levels possible for the MOF without actually assigning any LOF's to those levels. The introduction of a new MOF to the Dictionary File does not require the recreation of the Data File since no

LOF's can exist for that MOF. LOF's of such newly-introduced MOF's would have to be entered into the Data File by means of Data File maintenance.

Consider the MOF, (PGS), previously used as an illustrative example, with the following structure:



The MOF (PGS) could be properly constructed with the following dictionary cards:

MOF DESC 1 5	CODE 6	TEXTUAL SPELLING 7 72	STRUCTURE IND 73 74	75-80
(PGS)	2	PRIMATE GROUPS INVOLVED IN STUDY	3	
(PGS)		ANTHROPOIDEA	1	
(PGS)		CATARRHINI	2	
(PGS)		CATARHINI	$	
(PGS)		OLD WORLD ANTHROPOIDS	-	
(PGS)		PONGIDAE	3	
(PGS)		PONGIIDAE	$	
(PGS)		SIMIIDAE	-	
(PGS)		SIMIDAE	$	
(PGS)		GREAT APES	-	
(PGS)		CERCOPITHECIDAE	3	
(PGS)		PLATYRRHINI	2	
(PGS)		HAPALIDAE	3	
(PGS)		CEBIDAE	3	
(PGS)		PROSIMII	1	
(PGS)		LEMUROIDEA	2	
(PGS)		TASSIOIDEA	2	

The previous example (first listing) illustrates the Dictionary File records resulting from input of the above cards.

7.1.4.2 Dictionary Updating Since the Dictionary File is, by definition, open-ended, new LOF's may be added at any time. These LOF's may be incorporated into the Dictionary File either when the data points are encoded originally or when the system indicates that a previously undefined LOF has been encountered. In the latter case a pre-punched card containing the MOF designation and a clear text spelling of the undefined LOF will automatically be provided by the system. This will insure that the correct LOF is defined and that all such LOF's are considered. The LOF reference numbers will be provided in the dictionary printouts.

A previously undefined LOF can fall into any one of the following categories:

(1) Variant spelling for an existing LOF
(2) Synonym for an existing LOF
(3) New LOF

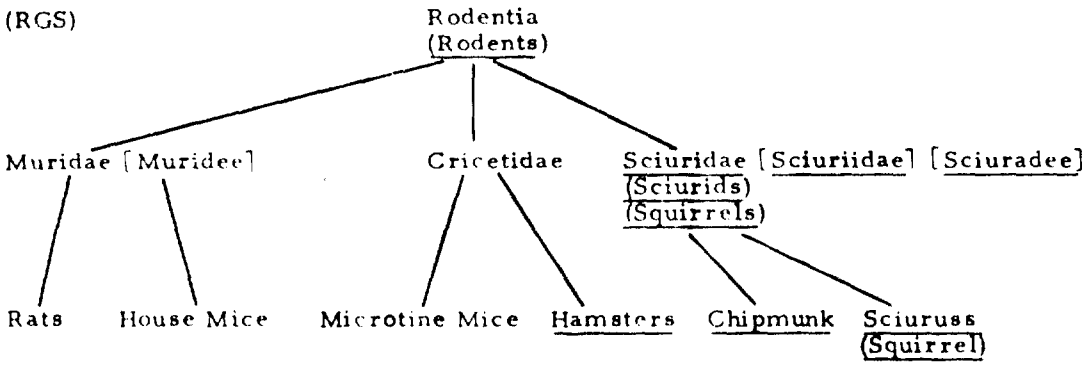
If the LOF is in category 1 or 2, it is incorporated into the Dictionary File merely by equating it to the appropriate dictionary entry. Card columns 75-80 are used to indicate the LOF reference number of the corresponding dictionary entry, and the structure indicator will contain the variant spelling ($) or synonym (=) symbol. If the LOF is a new word it must be related to an existing LOF unless the MOF has only one level. The relationship is determined by describing where in the tree structure the new LOF belongs, and is indicated by assigning a level in the structure indicator and by providing the LOF reference number of the (base) entry under which the new LOF should appear. Other new LOF's on the same tree branch are coded with a structure indicator, but without any LOF numbers. Thus the method for updating is identical to that for construction, with the exception that explicit LOF numbers must be included for certain entries. (These updating cards could be combined with the initial MOF construction cards to reconstruct any MOF's.)

Consider the MOF's, "Carnivore groups involved in study (CGS)" and "Rodent groups found during survey (RGS)", with the following structures, in which the underlined LOF's are to be added to the existing MOF's:

(CGS)

Canidae [Caniidae] Felidae (Cats) Ursiidae [Ursidae] (Bears)

(RGS)



The following dictionary cards would be required to include the preceding new LOF's (underlined) into the MOF's, (CGS), and (RGS), where N(k) equals the LOF reference number of the existing LOF entry k .

MOF DESC 1 5	CODE 6	TEXTUAL SPELLING 7 72	ST IND 73	75 80	Comments -
(CGS)		CANIIDAE	$	N(Canidae)	order is immaterial
(CGS)		CATS	=	N(Felidae)	
(CGS)		URSIIDAE	1		assignment of synonym permissive
(CGS)		URSIDAE	$		
(CGS)		BEARS	=		
(RGS)		RODENTS	=	N(Rodentia)	order is immaterial
(RGS)		MURIDEE	$	N(Muridee)	
(RGS)		HAMSTERS	3	N(Cricetidae)	
(RGS)		SCIURIDAE	2	N(Rodentia)	an order is essential
(RGS)		SCIURIIDAE	$		
(RGS)		SCIURADEE	$		
(RGS)		SCIURIDS	=		
(RGS)		SQUIRRELS	=		
(RGS)		CHIPMUNK	3		
(RGS)		SCIURUSS	3		
(RGS)		SQUIRREL	=		

The MOF, (PGS), could be properly constructed with the following dictionary cards:

MOF DESC 1 5	CODE 6	TEXTUAL SPELLING 7 72	STRUCTURE IND 73 74	75-80
(PGS)	2	PRIMATE GROUPS INVOLVED IN STUDY	3	
(PGS)		ANTHROPOIDEA	1	
(PGS)		CATARRHINI	2	
(PGS)		CATARHINI	$	
(PGS)		OLD WORLD ANTHROPOIDS	=	
(PGS)		PONGIDAE	3	
(PGS)		PONGIIDAE	$	
(PGS)		SIMIIDAE	=	
(PGS)		SIMIDAE	$	
(PGS)		GREAT APES	=	
(PGS)		CERCOPITHECIDAE	3	
(PGS)		PLATYRRHINI	2	
(PGS)		HAPALIDAE	3	
(PGS)		CEBIDAE	3	
(PGS)		PROSIMII	1	
(PGS)		LEMUROIDEA	2	
(PGS)		TASSIOIDEA	2	

The previous example (first listing) illustrates the Dictionary File records resulting from input of the above cards.

7.1.4.3 Dictionary Correction Due to the generic nature of many of the MOF's, structural changes in the Dictionary File would be difficult and cumbersome to accomplish -- and to describe -- in terms of updating. Moreover, such changes to the Dictionary File would make the Data File obsolete. For these reasons structural changes should be achieved by MOF reconstruction, limiting dictionary correction to such changes as would not affect entries in the Data File.

Correction of the dictionary is, therefore, restricted to the following functions:

(1) Change in textual description of a LOF or MOF -- indicated by a "/" in card column 73 (structural indicator).

(2) Delete any MOF or LOF entirely -- indicated by a "D" in card column 73.

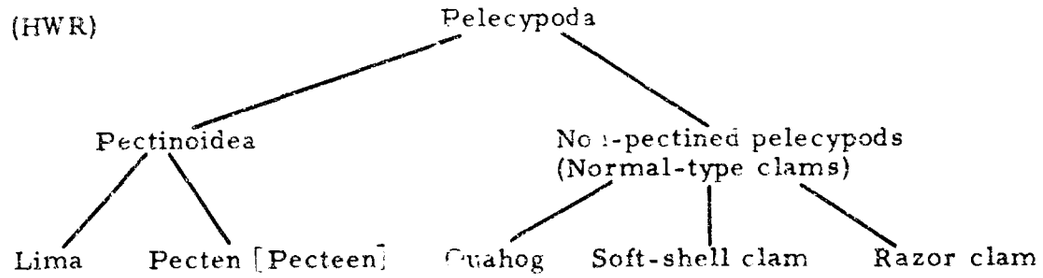
Correction of a variant spelling requires special consideration since such LOF's do not possess a unique LOF number by which they can be identified. Because of this the elimination of a variant spelling can never alter the structure of a MOF. Variant spellings can be physically removed from the Dictionary File if the structural indicator "D" is utilized with the textual description of the variant spelling. This is the only type of Dictionary File maintenance in which this description field, card columns 7-72, contains the LOF to be operated on. (For ease of processing the LOF number should also be indicated.) If a variant spelling is to be corrected it must be deleted and the correct variant spelling entered as an update.

Changes in the Dictionary File for LOF's (or MOF's) which are not variant spellings are accomplished by using the structural indicator "/" and their existing LOF number (or MOF designation). The deletion of such a LOF would normally be accomplished by use of "D" and its former LOF number; changing the textual description of a LOF to a blank field would also delete that LOF. In either event the LOF text would be considered u defined

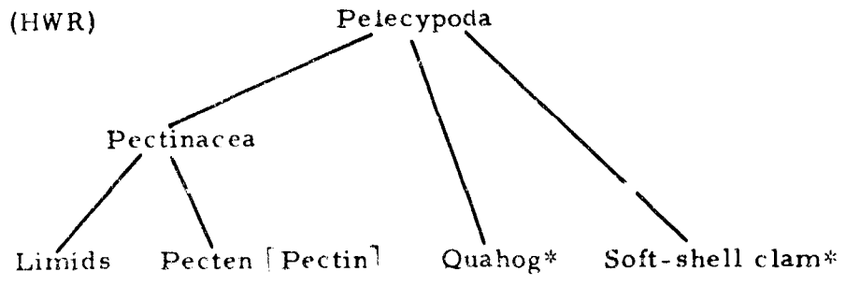
for input. In the latter case, however, the LOF reference number would not be undefined when used in the input data.

When a LOF is deleted, synonyms and lower level tree-structured LOF's remain in the Dictionary File unless explicitly deleted. Moreover, any future maintenance of the remaining LOF's must reference the original level of the LOF.

To illustrate the foregoing, consider the existing MOF, "Pelecypod groups found in water reservoirs (HWR)", with the following structure:



This MOF could be transformed into the MOF, "Clams found in drinking - water reservoirs (HWR), with the following structure:



Quahog and soft-shell clams are still physically at level 3 in the Dictionary File; however, logically they could be considered as being at level 2.

The above MOF transformation could be achieved by the following (correction) dictionary cards:

MOF DESC 1 5	CODE 6	TEXTUAL SPELLING 7 72	ST IND 73	 75 80
(HWR)	3	CLAMS FOUND IN DRINKING- WATER RESERVOIRS		
(HWR)		PECTINACEA	/	N(Pectinoidea)
(HWR)		LIMIDS	/	N(Lima)
(HWR)		PECTEEN	D	N(Pecten)
(HWR)		PECTIN	$	N(Pecten)
(HWR)			/	N(Non-pectinid p.)
(HWR)			/	N(normal-t.clams)
(HWR)			/	N(razor clams)

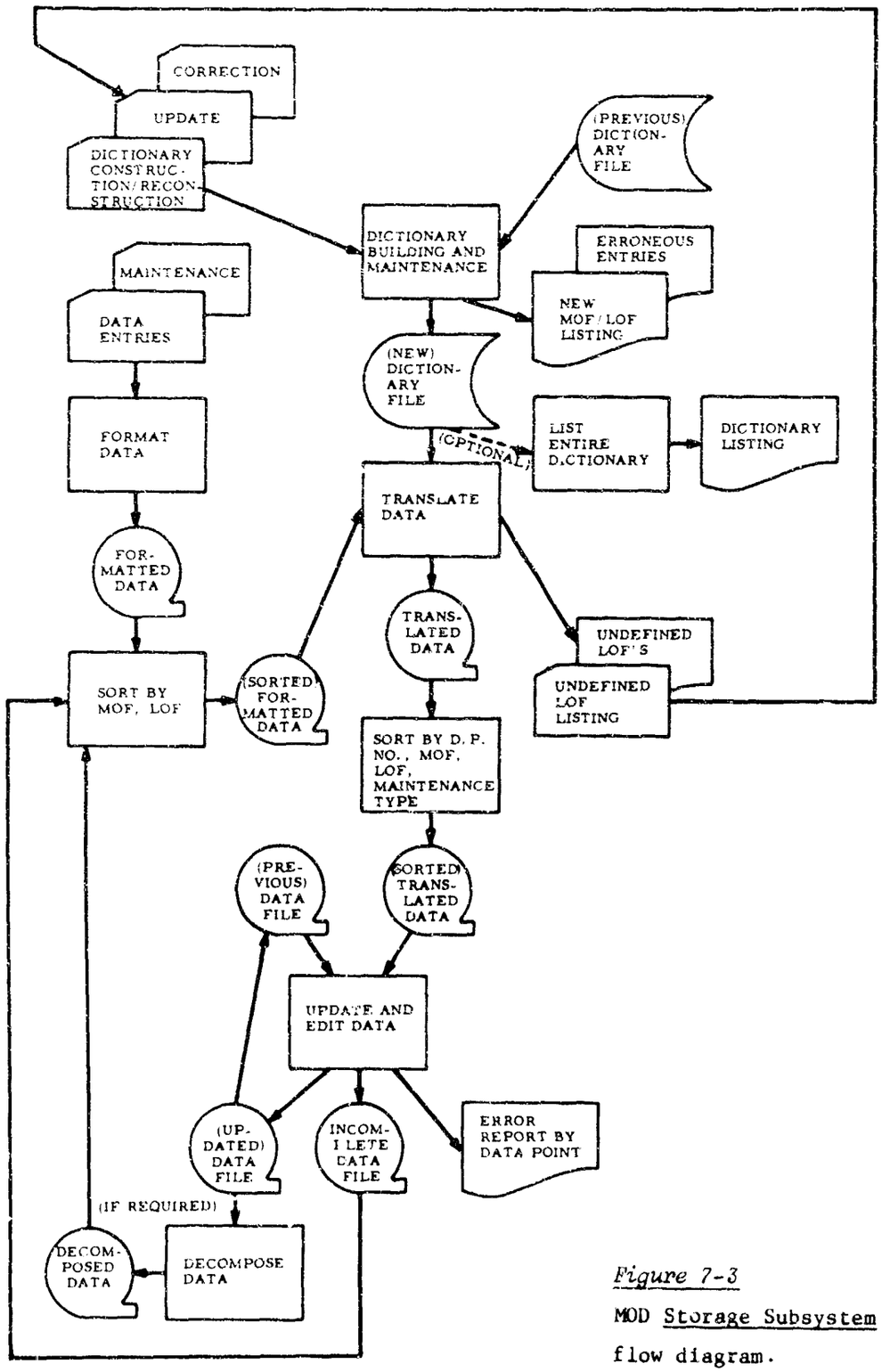


Figure 7-3

MOD Storage Subsystem flow diagram.

7.1.5 STORAGE PROCESSING

The objective of the MOD Storage Subsystem is to build and maintain a collection of valid data point records from the input data. This requires that the validity of incoming records be checked by the Dictionary File, hence the Dictionary File must also be built and maintained. The MOD Storage Subsystem has been designed so that the processing of the dictionary and the data can be accomplished either simultaneously or individually (see Fig. 7.3).

When both operations are to take place the dictionary processing is accomplished first. In an initial run it would be advantageous to predefine a subset of dictionary terms to reduce the number of undefined LOF's. In subsequent runs, maintenance to the dictionary might well include both previously undefined LOF's and newly defined LOF's for the current input data.

7.1.5.1 Dictionary Building and Maintenance All types of file maintenance input cards are processed to build or revise the Dictionary File. The original input sequence of these cards must be maintained since the order indicates the structural relationships of the LOF's within a MOF. Furthermore, this original sequence cannot be recreated by machine.

In general the type of maintenance to be performed is designated in columns 73-74 of the input cards, however, MOF reconstruction (or construction) can only be differentiated from regular updating by the presence of a level entry in the MOF maintenance card. MOF maintenance cards differ from LOF cards in that MOF cards contain a code in card column 6.

The sequence of processing in Construction, Reconstruction, and Updating of the Dictionary File is as follows:

(1) (TAPE ONLY) Build a magnetic tape record from input card that contains a generated serial number which is an ordered combination of the MOF order on the Dictionary File, any indicated LOF reference number, and the original input sequence number.

(2) (TAPE ONLY) Sort these generated records by their serial number. The textual spelling may be added as the minor field of the sort to facilitate processing corrections to variant spellings.

(3) Process input against existing numerical records in the Dictionary File.

(4) Build MOF record.

(5) Assign next sequential reference number to new LOF.

(6) Construct LOF code number.

(7) Build numeric order LOF record.

(8) (TAPE ONLY) Sort numeric order records into alphabetic order.

(9) (DISK ONLY) Determine index number of alphabetic order for numeric order and create alphabetic record.

(10) Print new dictionary entries alphabetically.

(11) List file maintenance errors, if any.

(12) At user's option, print the entire Dictionary File.

The dictionary listing have the following format:

(MOF short form)	MOF name or long form	Number of LOF entries
	LOF name	LOF reference number

Errors are listed separately after the dictionary listing and have the same general format as above plus an explanation of the MOF or LOF error, printed on the right side of the page.

This basic dictionary processing has the following variations, according to the type of maintenance being performed:

Construction or Reconstruction -- The previous numerical records for the MOF are disregarded. Any LOF's with designated reference numbers are processed first; the remaining LOF's are assigned reference numbers. MOF reconstruction requires that the Data File be regenerated to insure that the proper LOF code numbers are contained therein.

Update -- The previous numerical records for the MOF are retained and new records are generated as required.

Correction -- The previous numerical records are retained for all the LOF's in the MOF except for those which are deleted.

Since the order of the cards indicates the structure of the MOF, it must be assumed that the input order is correct; if these cards are not in the proper sequence the MOF will have to be reconstructed. Deletions, corrections, references to non-existent LOF's, and references to unidentifiable MOF's will be flagged as errors.

7.1.5.2 Data File Processing After an initial Dictionary File is built, input processing then creates the Data File from the data input cards. This processing not only creates new data point records but also corrects and updates existing data point records in the Data File. An existing Dictionary File is necessary in order to process the input MOF's and LOF's properly. Input entries are matched against the dictionary file to insure the validity of all MOF's and LOF's, also to convert qualitative LOF's to their numeric code number for internal storage in the Data File. Undefined LOF's, unallowable or invalid LOF's and MOF's, and any other detectable errors are listed during this procedure. Incomplete data point records are maintained as a separate incomplete Data File until some corrective action is taken. The input processing functions are performed by the following programs:

(1) FORMAT DATA -- This program transforms the free-form input data into fixed-format magnetic tape records in which each LOF is an individual

record. Each record contains the Data Point Number, the MOF designation, and one LOF. The LOF may be in the form of a LOF reference number, a textual spelling, or a numeric value. The FORMAT DATA program also assigns a code number to the MOF's that corresponds to the MOF order within the dictionary file. In addition, the various types of Data File maintenance cards are assigned a code in accordance with the order of file maintenance precedence. This maintenance code is used as a minor sort field after the data has been translated. Under the rules established for keypunching the input data, each input card is an entity, hence it can be processed in any desired order. It is not required that narrative (NAR) records be properly sequenced at this time.

(2) SORT FORMATTED DATA -- The output tape from the FORMAT DATA program is then sorted by the assigned MOF code number and by the LOF reference number (less MOF code digit and the checksum digit). At the same time, the Incomplete Data File is incorporated into the sort as a second reel of input. Both the new Formatted Data File and the Incomplete Data File will have the same format and may be considered as one entity during the succeeding processing. The purpose of this program is to speed the matching of MOF's and LOF's with the Dictionary File, however, if the Dictionary File is on magnetic tape, this operation becomes essential rather than merely a means of increased efficiency. This sort operation will sequence all the LOF's into alphabetic and numeric order within each MOF (but the LOF sequence has functional significance only for qualitative LOF's).

(3) TRANSLATE DATA - This program is the bridge between the input data and the Data File. Here, LOF records from the preceeding sort program will be compared with appropriate entries in the Dictionary File. Each qualitative LOF will be tested for definition. If defined, the LOF code number will be added to the LOF data record. LOF reference numbers will be matched against the numerical order section of the dictionary, in addition, the validity of their MOF code digit and checksum digit will also be determined (see Fig. 7.2). LOF textual spellings will be compared with the

alphabetic order section of the dictionary. If such a textual entry does not exist in the Dictionary File, the entire entry is listed and a card is punched. This card will contain the MOF designation and the textual spelling of the undefined LOF -- in the Dictionary File maintenance format -- so that it can be entered later into the Dictionary File maintenance programs without a need to create an entire entry and without danger of mis-punching the textual spelling. The punching of LOF cards will be summarized at the LOF level, e.g., even though several data points have the same undefined LOF for a given MOF, only one LOF card will be produced. Each quantitative LOF will be tested for validity as indicated by the MOF. Alternatively, this MOF validity indication may be added to the LOF record and tested in the edit program. The LOF record for any LOF which is undefined, or invalid, or which refers to an unidentifiable MOF will be flagged. But only the undefined qualitative LOF's will be listed at this time (allowing all undefined words to be analyzed with respect to the structure of their MOF). In consequence, this listing will be uncluttered and will correspond to the punched cards. All the errors will be listed later by data point number in the data editing program so that the errors can also be considered in terms of the entire data point.

(4) SORT TRANSLATED DATA -- The translated data, including all error records, will be sorted by Data Point Number, MOF designation, LOF, and file maintenance type. This will provide for the immediate updating and construction of the Data File from all of the appropriate LOF records. The narrative (NAR) records and those for any other non-retrievable "MOF" will be sequenced as the last records for each data point, and by continuation number, if applicable.

(5) UPDATE & EDIT DATA FILE -- The data point records of the Previous Data File will be updated by the output of the SORT TRANSLATED DATA program. For processing facility, sufficient main memory should be available to contain one data point record from the old Data File, all of the new LOF input records for a data point, and a buffer area for a new data point record. None of these data point records would have to include any narrative at this

time. One should have access to the entire set of new LOF's for a data point so that all of its component items can be written on the Incomplete Data File if necessary. If main memory availability is severely restricted, these incomplete data records could be purged by additional processing.

The updating and editing for each data point will consider one entire MOF at a time. Any necessary file maintenance operations will first be performed in accordance with the order of maintenance precedence. File maintenance errors such as deletion, addition, or replacement of non-existent entries will be listed. Then required edit checking will be done. In some instances, consistency among different MOF's may be tested. Several specific processing steps, necessitated by the characteristics of the MOD data, will also be carried out during the input processing for the Data File. For example, the data-reliability MOF "Computer evaluation of data point" will be calculated according to a suitable algorithm, and the resulting number stored as a numeric LOF for that MOF. Also, data points whose Specific Disease Agent is specified as a logical sum of positive and negative items will be split for storage and later processing into one point for all the positives (with a non-zero value) and one zero-valued data point for all the negative items. Finally, the entire newly formed data point will be searched to insure that all essential MOF's are present. Then record index of updated or new data point records will be appropriately revised.

An Incomplete Data File will be generated that will contain all of the LOF records for those data points which lack essential MOF's or have undefined qualitative LOF's. Such data points will not be included in the Updated Data File. Other types of LOF errors will merely cause that LOF to be eliminated from the appropriate file. If this elimination causes the loss of an essential MOF, the data point will be transferred to the Incomplete Data File.

Both the Updated Data File and the Incomplete Data File are in data point sequence. The purpose of the incomplete file is to simplify correction

requirements, also to facilitate the listing of these deficiencies until they are remedied. Hence, only new or updated data point records will be tested; unaltered records will simply be merged into the new (Updated) Data File. Existing data point records will not be eliminated from the Data File if erroneous corrections are made. Rather, these invalid corrections will be listed and then ignored in the processing. If one wished to purge these records onto the Incomplete Data File, the existing data point records would have to be decomposed into component LOF records.

(6) DECOMPOSE DATA FILE -- This operation is required if a MOF in the dictionary has been reconstructed. In this event the LOF code numbers in the data will usually be inaccurate and will have to be regenerated. This can be accomplished if every data point record in the Data File is decomposed into a group of separate LOF records. The LOF reference number can be determined as being the lowest level entry in the LOF code number contained in the Data File. (The regeneration of the Data File is possible because MOF reconstruction does not alter the reference numbers.) The entire Previous Data File can then be re-entered into the system in the form of LOF records, as additional input to the SORT FORMATTED DATA program. These separate LOF records will contain the MOF designation and the particular LOF item. For qualitative LOF's, this item will be the LOF reference number.

7.2 RETRIEVAL SUBSYSTEM

The function of the MOD Storage Subsystem is to create a data base from which the desired MOD output results can be produced. The MOD user will obtain this output by means of a query to the MOD system. His query must describe the following three aspects of the desired output:

(1) Retrieval conditions -- any characteristics that the data must contain in order to be considered for output (such as specific disease agent, or species infected, or time period).

continued next page

(2) Synthetic or manipulative operations -- any operations which must be performed upon the retrieved data prior to output (such as combining points by averaging their values).

(3) Output specifications -- the type, format, and content of the desired output (such as map projection to be used).

These are accomplished in sequence given since a subset of data points must first be considered, then operated upon, and, finally, displayed.

The Retrieval Subsystem (shown in Fig. 7.4) will now be discussed in detail as it is more related, logically and physically, to the Storage Subsystem than to the other two subsystems. Moreover, retrieval is the first (and most fully developed) aspect of the entire query procedure. Manipulative operations and output specifications are independent of retrieval, and both of these will be discussed later.

7.2.1 RETRIEVAL LANGUAGE

In any retrieval system, items are selected for retrieval which satisfy the given (query) conditions. The manner in which these conditions are expressed is of the utmost importance for effective retrieval. At the present stage of development of the MOD system there has been insufficient experience in the areas of retrieval usage to determine optimal specifications appropriate to the requirements (and background) of the potential bio-medical users. For this reason an interim retrieval language has been established. Based upon experience gained in actual use of the MOD system, the interim retrieval language can be modified to yield a more elaborate -- and efficient -- "ultimate" retrieval language. But with this present MOD system design, the specific retrieval request would be formulated by the data analyst from a more generalized query made by the bio-medically oriented user. (Of course the user himself could formulate the retrieval request if he were confident that he understood fully all the logical facets of his

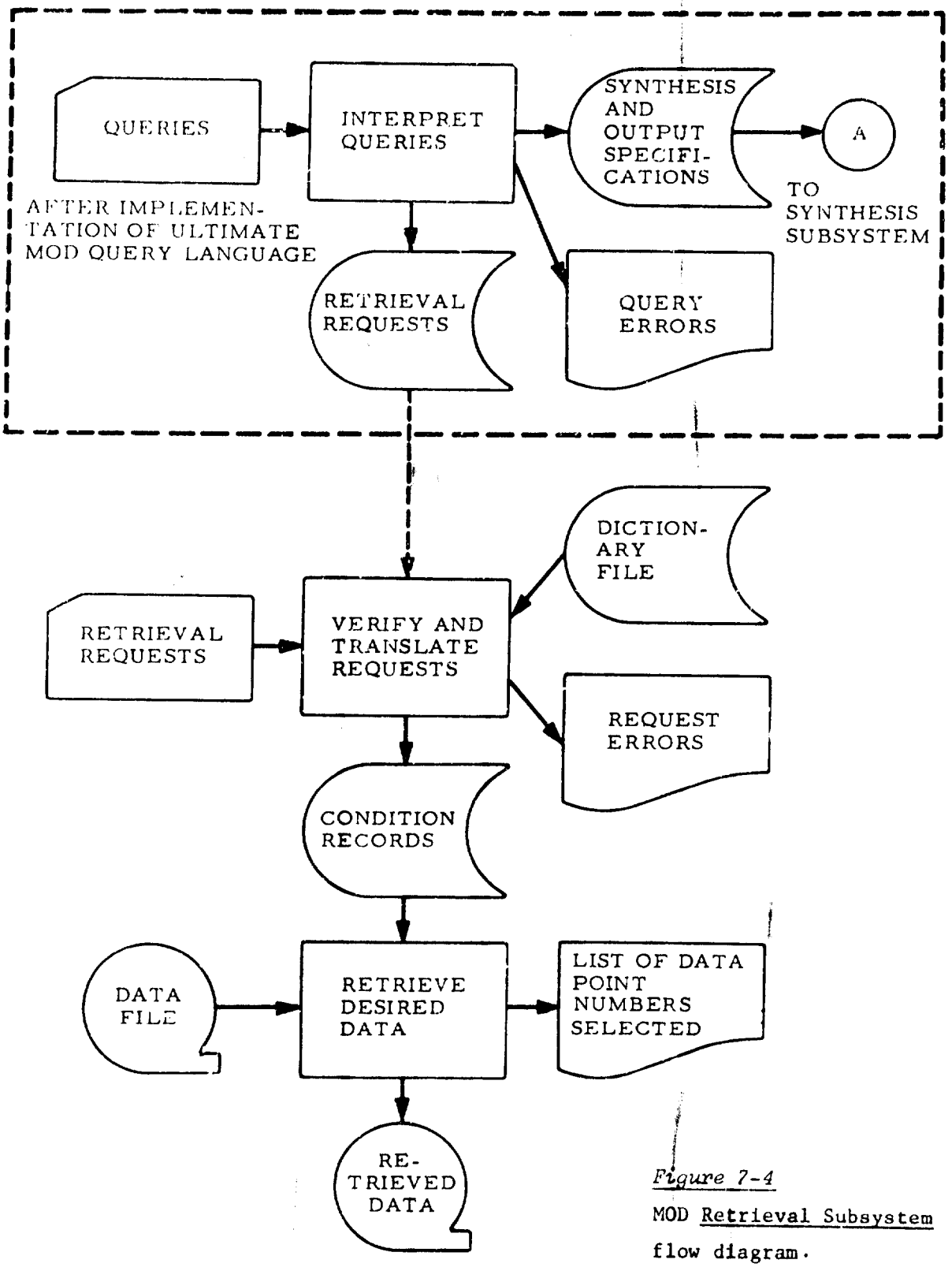


Figure 7-4

MOD Retrieval Subsystem flow diagram.

query. When there has been sufficient experience with the implemented MOD system, the ultimate MOD retrieval language can also be formulated.

From a system viewpoint, those portions of the MOD query which relate to retrieval can be expressed in terms of the interim language by a preliminary processing program, then be operated upon by the Retrieval Subsystem as specified in this section. To provide for this transition and to allow for all logical requests, the interim language consists of all the basic logic functions and operations of a general retriefal system, expressed in the most direct and concise manner.

Let us consider first the rules of logic which pertain to the use of operators to connect conditions. In the following, A, B, C, and D each represents any retrieval condition.

The logical operators, AND or OR, can combine any pair of conditions, and the result is itself a condition. In the MOD interim retrieval language "+" and "/" will be used to indicate AND and OR respectively. The meaning of these operators are:

/ (OR) The condition A / B is satisfied if A is true, or if B is true, hence, also, if A and B are both true.

\+ (AND) The condition A + B is satisfied if, and only if, both A and B are true.

Since the result of a logical operation upon two conditions is itself a condition, another condition can be combined with it. But these combinations are not associative, hence parenthesis must be used to indicate the meaning of certain combinations.

If the logical operators are the same, parenthetical grouping is unnecessary.

Example: A / B / C = (A / B) / C = A / (B / C)

A + B + C = (A + B) + C = A + (B + C)

If the logical operators are mixed, parenthetical grouping is essential for proper meaning.

Example:

$$A / (B + C) = (A/B) + (A/C) \neq (A + C) / (B + C) = (A/B) + C$$

$$A + (B/C) = (A + B) / (A + C) \neq (A/C) + (B/C) = (A + B) / C$$

Since parenthetical grouping is unnecessary for similar operations, more than one condition can appear within a parenthesis, e.g.

$$(A / B / C) + D \qquad\qquad (A + B + C) / D$$

Reapplying these rules, an infinite number of levels of parenthetical grouping can be established. However, any expression which contains higher levels of grouping can be reduced to one level of parenthetical grouping by appropriate repetition.

Example: $A / (B + (C / D)) = A / (B + C) / (B + D)$

$A + (B / (C + D)) = A + (B / C) + (B / D)$

$((A + B) / C) + D = (A + B + D) / (C + D)$

Thus the interim retrieval language can perform any desired retrieval operation, if the following two rules are followed:

(1) Parenthesis are only used where necessary (between unlike logical operators but not between like operators).

(2) Only one level of parenthesis is allowed (higher levels must be manually reduced).

Thus far we have considered conditions abstractly, and treated each condition as an entity. These conditions do actually apply to MOD data however, and consist of several components. These components establish a criterion which will either be true or false for every data point record of the Data File. We are not merely searching for the presence of an item in the Data File; it is necessary that this item be considered within the proper context, i.e., a specific LOF within a particular MOF. For added

flexibility we can allow the relationship of the LOF to the MOF to be other than equality.

The three components of a retrieval condition are:

(1) MOF designation

(2) LOF description

(3) Relational operator

The relational operators defined for the MOD retrieval system are:

= Equality (including synonyms, variant spellings, and less generic tree structured components).

≡ Identity (including variant spellings, but not synonyms).

≠ Inequality (i.e., not equal to).

< Less than (significant only for numeric values).

> Greater than (significant only for numeric values).

The usual MOF designations are used without their parentheses in a retrieval request because the existence of the relational operators easily distinguishes MOF's from LOF's. Moreover, since one level of parenthetical grouping is allowed for logical grouping, use of parenthesis for other purposes in the language should be avoided.

To be consistent with our rule that only one level of parenthetical grouping be allowed in a retrieval request, each condition is to contain one and only one MOF and one LOF. If a criterion logically includes two possible LOF's for a MOF, the MOF must be explicitly stated twice with the proper logical operators.

The LOF description of quantitative LOF's will consist of their actual numeric value. Qualitative LOF's can be described either by their textual spelling or their reference number (but textual description would probably be the more useful method of specifying a qualitative LOF).

7.2.2 RETRIEVAL REQUEST CARDS

Retrieval requests will be activated in the MOD system by means of retrieval request input cards. Several different sets of criteria may be requested at the same time, and these will be distinguished by being assigned a different question number.

The retrieval cards consist of two fields. The first contains identification data, question number (and, possibly, continuation card number). The remainder of the card contains the requests in free form with non-essential blanks optional. Of course these requests must be formulated in accordance with all of the rules described in the preceding section. Identification data will be listed on all output retrieval reports.

The maximum number of conditions per question and the maximum number of questions that can be processed at the same time will have to be determined prior to establishing precise rules for these items.

A sample set of request cards would appear as follows:

```
JDHS 5/12/67    1    (MFX = GOOD / QXF = HIGH) + VAL > 25.3
JDHS 5/12/67    2    (MFX = GOOD / MFX = FAIR) + (VAL > 20
JDHS 5/12/67    2    / PVL > .05) + SDA ≠ L. POMONA
JDHS 5/12/67    3    PHD ≠ WILD
```

7.2.3 RETRIEVAL PROCESSING

The retrieval subsystem reads the request cards, checks the validity of their form and content, and obtains any required LOF code numbers. This subsystem then tests each data point record in the Data File on a match/synonym basis and writes the selected records onto one or more magnetic tape files.

First, validity of the format of the retrieval request cards is tested; detected errors will be listed. If there exists an extraneous parenthesis

in the request, the "corrected" interpretation will be listed as a flag and the requests processed in accordance with this interpretation.

After the validity of the format of the entire retrieval request has been established, a condition record is generated for each condition in the request. These condition records contain the following elements:

(1) Question number.
(2) Condition number.
(3) Designated MOF.
(4) Specified LOF.
(5) Required relational operator.
(6) Next operation if condition is true.
(7) Next operation if condition is false.

The question number is obtained directly from the request cards. The condition number indicates the sequence of each condition within a question. The logical sequence must be maintained in order to execute the retrieval processing properly.

After these operations the validity of the requested MOF's and LOF's is determined. For this purpose the conditions must be considered first in MOF, then in LOF order. The volume of these requests will probably be such that an external sort of the conditions will be unnecessary. The Dictionary File is used to determine the validity of the MOF's and LOF's. The LOF element of the condition records for quantitative LOF's will contain the requested value. For qualitative LOF's, this element will contain those portions of the LOF code number which are appropriate to the request. Generally, this consists of all the reference number components of the code number down to the level of the LOF being considered. (The level can be determined from the Dictionary File by the presence of the first zero reference code or the last reference code within the LOF code number.) But if the desired relationship is one of identity, the entire LOF code number is placed in the LOF element of the condition record. The relational operator of the

condition is tested to insure that it is appropriate for the type of LOF being considered. This operator is then placed in the relation element of the condition record. (The internal indication for identity can be the same as that for equality since the composition of the code number will distinguish between these relationships.)

The contents of the "next operation" field can be determined from the logical operator (which immediately follows the present condition) if, as is the case, there is only one level of parenthetical grouping and if the logical sequence of the conditions is preserved:

NEXT LOGICAL OPERATOR	NEXT OPERATION IF PRESENT CONDITION IS:	
	TRUE	FALSE
None	Select	Reject
/ outside of parenthesis	Select	Test next condition
/ inside of parenthesis	Test next condition outside of parenthesis or select if none exists.	Test next condition
+ outside of parenthesis	Test next condition	Reject
+ inside of parenthesis	Test next condition	Test next condition outside of parenthesis or reject if none exists.

This selection or rejection refers to the entire data point record being tested. The non-existence of a next logical operator is considered within the format of the present question if more than one question has been requested.

The condition records for some fundamental types of requests are now provided. In the following, the MOF, LOF, and relational operator components of a condition have been represented by a single letter for simplicity, and each request has been assigned a different question number.

QUESTION CONDITION	CONDITION NUMBER			NEXT OPERATION: IF TRUE - IF FALSE	
	QUESTION NUMBER		CONDITION RECORD		
A/B	1	1	A	Select	to 2
	1	2	B	Select	Reject
A+B	2	1	A	to 2	Reject
	2	2	B	Select	Reject
A+(B/C)	3	1	A	to 2	Reject
	3	2	B	Select	to 3
	3	3	C	Select	Reject
(A+B)/C	4	1	A	to 2	to 3
	4	2	B	Select	to 3
	4	3	C	Select	Reject
A/(B+C)/D	5	1	A	Select	to 2
	5	2	B	to 3	to 4
	5	3	C	Select	to 4
	5	4	D	Select	Reject
(A+B)/(C+D)	6	1	A	to 2	to 3
	6	2	B	Select	to 3
	6	3	C	to 4	Reject
	6	4	D	Select	Reject

Any errors detected in the format or content of a retrieval question will cause that question not to be processed. After all questions and conditions have been verified, the user will have an option as to whether or not retrieval questions without errors should be processed if other questions in his request contain errors.

After all the condition records have been generated for a request, each data point in the Data File is tested against this set of condition records by comparing the MOF, LOF, and relational operator. The location of the LOF's within each data point record is indicated in the data record index.

The LOF description in the condition records will determine the matches. Every data point is compared with the retrieval conditions set forth by each retrieval question. The methods employed in the Data and Dictionary Files have been established so that synonyms, variant spellings, and tree-structure relationships do not have to be handled by long retrieval lists. Moreover, the condition records enable the retrieval processing to test only those LOF's required to ascertain whether a data point should be selected or rejected.

If a data point record is selected it is written onto a magnetic tape file and listed by Data Point Number as having been retrieved. Various questions may be output onto different tape units, alternatively, the question number may be appended to the data point records selected by that question. The formats of the Updated Data File and the Retrieved Data File are identical, hence either file may be used for subsequent output processing. Thus synthetic, or manipulative operations, or output specifications may also work against the entire Data File.

7.2.4 ALTERNATE LOF CODING PROCEDURE

A unique feature of the preceding Storage and Retrieval Subsystems is the method of coding the qualitative LOF's. The code number of each qualitative LOF is constructed to indicate the structural relationship for retrieval purposes. The Data File contains those code numbers that consist of a series of numbers whose total length is one greater than the number of levels within the MOF. The Retrieval System scans all or part of this series of numbers to determine if a selection criterion has been satisfied.

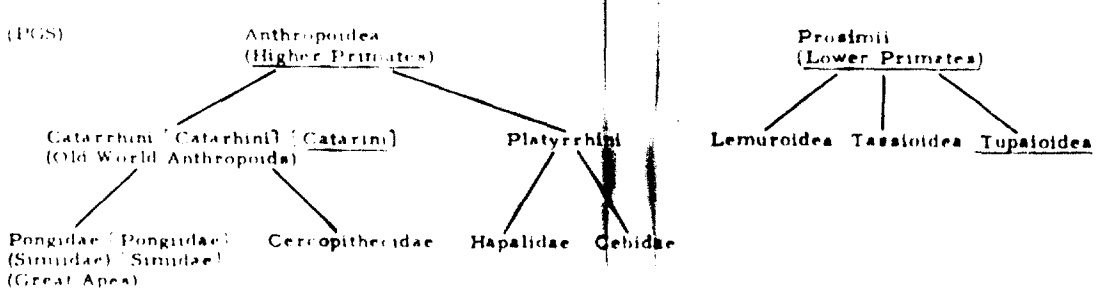
A LOF code number that consisted of only one number would suffice for retrieval purposes if that number were properly formulated, and it is with this consideration that we present the following alternative procedure.

Each LOF would have an original reference number. The LOF would then be sorted by its descending structure relationships within the MOF. This sequence would determine the present code number for the LOF's. A range of code numbers consisting of the first and the last code numbers which represent its structural relationship for each LOF could then be ascertained. This method is possible because the range of code numbers has no missing members in it -- because of the structural sequence.

This single number code system would substantially shorten the lengths of the Data and Dictionary Files and would make the retrieval process more direct. Each LOF entry in the Data File would consist of only two numbers, the original reference number and the present code number. The code number in the dictionary would be reduced to two numbers for all qualitative MOF's first and last range value. The range values would be used for normal retrieval and the present code numbers for "synonym lock-out" retrieval. If a LOF number were in the range of the requested LOF, the criterion would be satisfied unless synonym lock-out were desired, in which case only an exact match would suffice.

Weighing the pros and cons of this alternative method, the advantages of brevity seem to be more than offset by the requirement that the entire Data File (rather than just the new or incomplete data points) be translated against the Dictionary File after virtually any type of file maintenance is performed on the Dictionary File. (This alternative method would not affect the remainder of the MOD system.)

Consider again the MOF, "Primate groups, involved in study (PGS)", with the tree structure shown, underlined LOF's to be added to the MOF:



The MOF, (PGS), would appear in the Dictionary File as follows -- after its initial construction (without the underlined entries). The disk format is used for brevity.

NAME	CODE	RANGE	REF
ANTHROPOIDEA	1	1 10	1
CATARHINI	2	2 7	2
CATARRHINI	2	2 7	2
CEBIDAE	10	10 10	10
CERCOPITHECIDAE	7	7 7	7
GREAT APES	6	4 6	6
HAPALIDAE	9	9 9	9
LEMUROIDEA	12	12 12	12
OLD WORLD ANTHROPOIDS	3	2 7	3
PLATYRRHINI	8	8 10	8
PONGIDAE	4	4 6	4
PONGIIDAE	4	4 6	4
PROSIMI	11	11 13	11
SIMIDAE	5	4 6	5
SIMIIDAE	5	4 6	5
TASSIOIDEA	13	13 13	13
1	1	1 10	1
3	2	2 7	2
9	3	2 7	3
11	4	4 6	4
15	5	4 6	5
6	6	4 6	6
5	7	7 7	7
10	8	8 10	8
7	9	9 9	9
4	10	10 10	10
13	11	11 13	11
8	12	12 12	12
16	13	13 13	13

Note that the reference number and code number are identical -- after the initial construction of a MOF.

After the new (underlined) LOF's are added, (PGS) would be as follows:

NAME	CODE	RANGE	REF
ANTHROPOIDEA	1	1 11	1
CATARHINI	3	3 8	2
CATARINI	3	3 8	2
CATARRHINI	3	3 8	2
CEBIDAE	11	11 11	10
CERCOPITHECIDAE	8	8 8	7
GREAT APES	7	5 7	6
HAPALIDAE	10	10 10	9
HIGHER PRIMATES	2	1 11	14
LEMUROIDEA	14	14 14	12
LOWER PRIMATES	13	12 16	15
OLD WORLD ANTHROPOIDS	4	3 8	3
PLATYRRHINI	9	9 11	8
PONGIDAE	5	5 7	4
PONGIIDAE	7	5 7	4
PROSIMI	12	12 16	11
SIMIDAE	6	5 7	5
SIMIIDAE	6	5 7	5
TARSIOIDEA	15	15 15	13
TUPAIOIDEA	16	16 16	16
1	1	1 11	1
4	3	3 8	2
12	4	3 8	3
14	5	5 7	4
16	6	5 7	5
7	7	5 7	6
6	8	8 8	7
13	9	9 11	8
8	10	10 10	9
5	11	11 11	10
16	12	12 16	11
10	14	14 14	12
19	15	15 15	13
9	2	1 11	14
11	13	12 16	15
20	16	16 16	16

Note that the LOF code numbers are <u>not</u> now the same as the LOF reference numbers.

7.3 SYNTHESIS SUBSYSTEM

After all the pertinent data point records have been selected by the Retrieval Subsystem the Synthesis Subsystem performs necessary and desirable refining operations upon these points. The synthesized data is then used to produce maps and reports. These refinements consist of both necessary synthesizing operations, which are required to combine the data properly, and optional manipulative calculations, which are specified by the MOD user in his query request. Operation of this subsystem is diagrammed in Fig. 7.5.

Since geographic considerations are of the utmost importance throughout the MOD system, the geographic location of each data point must be adequately represented to fulfill all functional requirements. These requirements include:

(1) Validation of input data -- specified in terms of political units, or longitude and latitude, or both.

(2) Consistent internal storage of the location in the MOD Data File.

(3) Convertibility to either verbal descriptions (for output reports) or to X, Y coordinates (for mapping).

(4) Proper interpretation in query requests.

(5) Combination or coordination characteristics by which the data points can be combined, refined, and enhanced for output representation.

7.3.1 DICTIONARY FILE (LOCATION FUNCTIONS)

In the MOD system the Dictionary File is required to accomplish the following two functions dealing with geographic locations:

(1) Gazetteer function -- in which all geographic names must be described in terms of a generic tree-structure with synonyms and variant spellings (like the MOF's previously discussed).

continued next page

(2) Grid function -- in which a sufficient number of geographic points are identified to provide mapping coordinates; these coordinates must also be associated with some geographic name and must fit into the same logical form as do the other entries in the Dictionary File.

The gazetteer function can be achieved if the geographic names are described in terms of their political unit designations. These designations are mutually exclusive and provide a tree-structured hierarchy of country, province/state, county, and smaller unit. The smaller unit could consist of cities, towns, military installations, etc. Additional geographic levels may be added for continent, area of a country, parts of a state, etc. The gazetteer function allows construction of regional areas from any group of political units which are of the same tree level, e.g., the countries which comprise Southeast Asia, the states which make up the southwest portion of the United States, and the counties which constitute southern California.

For the proper operation of the Dictionary File all entries in a given MOF must have the same number of tree-structure levels, but it is not required that all of these have positive values; if no appropriate group name can be assigned for a particular collection of political units, the group designation is left blank. In some instances several geographic area levels may be constructed by nesting of mutually exclusive political units. With the proposed system it is also possible to construct geographic areas from a subset of political units which are not mutually exclusive with a higher level of political unit, for example, the "Delmarva peninsula" or "Rocky Mountains." From the viewpoint of tree structure, the level of such an entry would be both higher and lower than the state level. Actually, as will be shown, such a data point would be assigned to a coordinate within one of the appropriate states for mapping purposes. For this reason there would be advantage in having such terms as "Delmarva" undefined in the Dictionary File, allowing the data analyst to re-designate such a data point after it was rejected in the MOD Storage Subsystem. Possible re-designations for

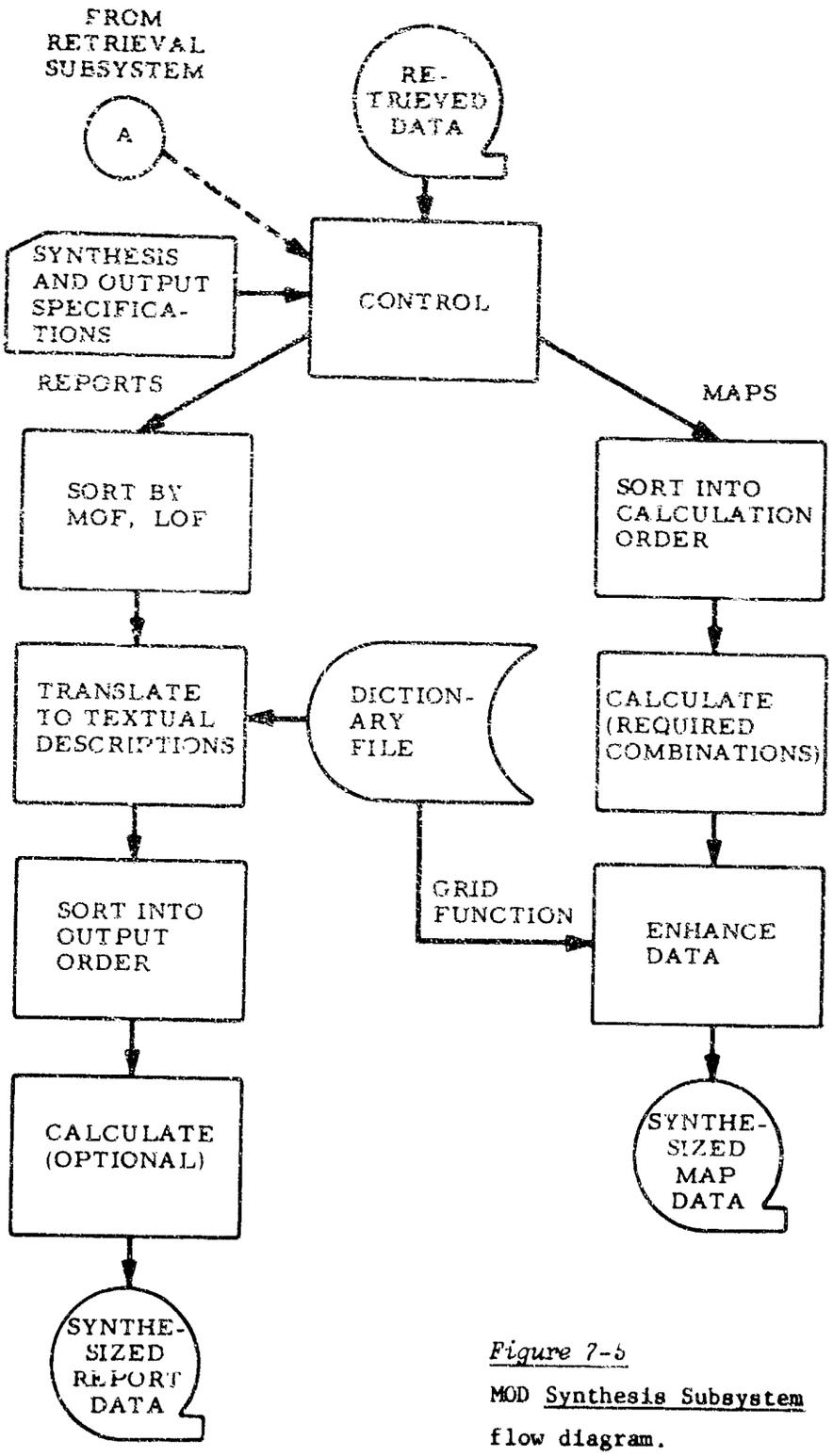


Figure 7-5

MOD Synthesis Subsystem flow diagram.

"Delmarva" would include "Maryland, Eastern Shore", "Virginia, Eastern Shore", and "Delaware". The geographic area, "Eastern Shore", would require also a state designation in order to find the proper dictionary-type entry.

Certainly the Dictionary File will contain many non-unique geographic names. For example, there are many "Washington" counties and several "Washington" cities in the United States, and such names must have additional geographic designations in order to make the entry unique. Ordinarily the geographic location of input data will include more than one designation, e.g., city, state, country, etc.

Geographic input names can be processed as follows in the MOD Storage Subsystem:

(1) Each geographic designation is treated as a LOF; LOC itself is treated as one MOF.

(2) In the TRANSLATE program all dictionary entries which match the alphabetical spelling of each such LOF are carried on LOF records for the data point.

(3) In the last pass of the SORT TRANSLATED DATA program all LOF records for the location of each data point are matched.

Since the LOF code number for each location contains the LOF reference numbers for all the higher-level geographic names, only one LOF code number will be consistent with the other higher-level reference numbers, assuming that the geographic location was sufficiently specified in the input. By this method the geographic locations need be specified by only enough levels to be uniquely defined. Obviously, conventions would have to be defined so that, for example, the single input location entry, NEW YORK, would always be interpreted to mean NEW YORK STATE. If the location designation(s) were insufficient for unique interpretation, the location entry would be listed as an error. If desired, the several possible locations for that entry could also be provided.

The grid function of the Dictionary File can be accomplished if the longitude and latitude coordinates that are to be used in mapping are assigned for the center-of-area of each defined geographic name. Then additional point locations must be associated with the center-of-area point and with the geographic name, to represent the geographic extent of the area. These additional locations (which may be called grid points) are required for shaded maps; they can also be used in contour maps. Grid points are determined in relationship to a grid size (the coordinate distance between grid points in a given geographic area). It is not essential that all areas of the earth be given the same grid size. Large bodies of water, deserts, and so forth should also be assigned coordinates so that such areas can be recognized and differentiated from areas likely to have valid contributory disease/environmental data. This type of distinction is essential in areas in which the gird size is large (i.e., coarse). The designation of non-applicable areas (i.e., those without valid contributory data) may be used to enhance all types of mapped output since they allow differentiation between non-applicable locations and locations for which no (retrieved) data points exist. This facility is considered essential for computer production of contour maps.

The gazetteer function of the Dictionary File can be fulfilled in exactly the same manner as has been described for the dictionary functions, with the exception that additional processing is required to combine the various LOF level designations into one unique location. If the longitude and latitude coordinate values are used for the LOF reference number, rather than somewhat randomly created integers, these values provide a basis for the gird function of the Dictionary File. A prime characteristic of the LOF reference numbers is that each one must be unique within the MOF. This unique characteristic of longitude and latitude coordinate points can be realized if each higher geographic level in the MOF is assigned an additional digit that is not significant from the standpoint of geographic location. Thus, if the finest grid considered is $1/10^{0}$ for the lowest possible level

of location designation, the next higher level could be designated in terms of $1/100^{\circ}$, where the least significant digit is not zero. These higher level reference numbers would be the approximate center point of the geographic area defined. The lowest geographic level coordinates of each grid point would have to be manually assigned. The higher level coordinates for those points could then either be manually assigned or computer generated. The latter operation is possible because the grid function can be considered as a tree structure in which each level is completely described by the next lower level.

There need be and should be only one building and maintenance program to satisfy both gazetteer and grid functions of the Dictionary File.

A single program would insure that all the location data were consistent. It seems desirable, however, that the gazetteer and grid entries exist in two different physical files in the MOD system in order to facilitate their use. One reason for this is that the additional grid points which are defined to indicate the geographic area associated with a location name are unnecessary for the lowest designation in the Dictionary File. Using separate files, the gazetteer records would be maintained in both alphabetic and numeric order, as would comparable records in the Dictionary File. Grid-point records, perhaps stored as a separate grid file, would be sequenced by the grid coordinates of all the component locations; they need not include the synonym and variant spelling entries. All bodies of water, etc., could be grouped under one tree level, and appear only in this grid file. If desired, there could be a dictionary of body-of-water names. These names could form a tree structure in themselves with synonyms and variant spellings for the input and retrieval of data concerning quatic environments. But such a structure would have to be a separate branch of the location tree-structure since the geographic locations of water bodies cannot always be uniquely correlated with the MOD political unit boundaries.

The building and maintenance operations for the gazetteer and grid functions will be accomplished in a manner similar to that for any MOF in the MOD Dictionary File, although additional processing is required to create entries appropriate for these functions.

The gazetteer and grid cards will have the same format and contents as ordinary Dictionary File cards, with the following exceptions:

(1) Grid cards for a location will have no textual description.

(2) Each lowest location level (on both gazetteer and grid cards) must contain a reference number which consists of its longitude and latitude.

(3) The format of the input cards will have a shorter field for the textual description, the relational-indication field will be moved to the left, and the reference-number field will be longer to accomodate exception (2).

For example, location input cards and the resultant Dictionary File records for the State of Delaware might appear as shown on the next three pages. (The grid size has been selected as $1/5^{C}$, and higher geographic levels have been omitted for brevity.)

Dictionary File

Textual Description	Relational Indicator	Reference Number	
		Long.	Lat.
DELAWARE	1		
NEW CASTLE	2		
WILMINGTON	3	-76.6	+39.8
ELKTON	3	-76.8	+39.6
	3	-76.8	+39.8
	3	-76.8	+39.4
	3	-76.6	+39.4
KENT	2		
	3	-76.8	+39.2
DOVER	3	-76.6	+39.2
	3	-76.6	+39.0
	3	-76.4	+39.0
SUSSEX	2		
OWENS TRACT STATE FOREST	3	-76.6	+38.8
OWENS FOREST	$		
OWENS TRACT	$		
ELLENDALE STATE FOREST	=		
ELLENDALE FOREST	$		
MILTON	3	-76.4	+38.8
	3	-76.6	+38.6
	3	-76.4	+38.6
	3	-76.2	+38.6

These cards would result in the following gazetteer entries in the Dictionary File.

Name (Textual Descr.)	Code Number						Reference Number	
DELAWARE	-76.581	+39.143	0	0	0	0	-76.581	+39.143
DOVER	-76.581	+39.143	-76.61	+39.11	-76.6	+39.2	-76.6	+39.2
ELKTON	-76.581	+39.143	-76.72	+39.61	-76.8	+39.6	-76.8	+39.6
ELLENDALE FOREST	-76.581	+39.143	-76.41	+38.71	-76.6	+38.8	-76.6001	+38.8	-76.6001	+38.8
ELLENDALE STATE FOREST	-76.581	+39.143	-76.41	+38.71	-76.6	+38.8	-76.6001	+38.8	-76.6001	+38.8
KENT	-76.581	+39.143	-76.61	+39.11	0	0	-76.61	+39.11
MILTON	-76.581	+39.143	-76.41	+38.71	-76.4	+38.8	-76.4	+38.8
NEW CASTLE	-76.58_	+39.143	-76.72	+39.61	0	0	-76.72	+39.61
OWENS FOREST	-76.581	+39.143	-76.41	+38.71	-76.6	+38.8	-76.6	+38.8
OWENS TRACT	-76.581	+39.143	-76.41	+38.71	-76.6	+38.8	-76.6	+38.8
OWENS TRACT STATE FOREST	-76.581	+39.143	-76.41	+38.71	-76.6	+38.8	-76.6	+38.8
SUSSEX	-76.581	+39.143	-76.41	+38.71	0	0	-76.41	+38.71
WILMINGTON	-76.581	+39.143	-76.72	+39.61	-76.6	+39.8	-76.6	+39.8
ELKTON	-76.581	+39.143	-76.72	+39.61	-76.8	+39.6	-76.8	+39.6
CASTLE	-76.581	+39.143	-76.72	+39.61	0	0	-76.72	+39.61
KENT	-76.581	+39.143	-76.61	+39.11	0	0	-76.61	+39.11
ELLENDALE STATE FOREST	-76.581	+39.143	-76.41	+38.71	-76.6	+38.8	-76.6001	+38.8	-76.6001	+38.8
OWENS TRACT STATE FOREST	-76.581	+39.143	-76.41	+38.71	-76.6	+38.8	-76.6	+38.8
DOVER	-76.581	+39.143	-76.61	+39.11	-76.6	+39.2	-76.6	+39.2
WILMINGTON	-76.581	+39.143	-76.72	+39.61	-76.6	+39.8	-76.6	+39.8
DELAWARE	-76.581	+39.143	0	0	0	0	-76.581	+39.143
SUSSEX	-76.581	+39.143	-76.41	+38.71	0	0	-76.41	+38.71
MILTON	-76.581	+39.143	-76.41	+38.71	-76.4	+38.8	-76.4	+38.8

These cards would also result in the following grid entries in the Dictionary File.

Name	Code Number					
DELAWARE	-76.581	+39.143	0	0	0	0
SUSSEX	-76.581	+39.143	-76.41	+38.71	0	0
	-76.581	+39.143	-76.41	+38.71	-76.2	+38.6
	-76.581	+39.143	-76.41	+38.71	-76.4	+38.6
MILTON	-76.581	+39.143	-76.41	+38.71	-76.4	+38.8
	-76.581	+39.143	-76.41	+38.71	-76.6	+38.6
OWENS TRACT STATE FOREST	-76.581	+39.143	-76.41	+38.71	-76.6	+38.8
KENT	-76.581	+39.143	-76.61	+39.11	0	0
	-76.581	+39.143	-76.61	+39.11	-76.4	+39.0
	-76.581	+39.143	-76.61	+39.11	-76.6	+39.0
DOVER	-76.581	+39.143	-76.61	+39.11	-76.6	+39.2
	-76.581	+39.143	-76.61	+39.11	-76.8	+39.2
NEW CASTLE	-76.581	+39.143	-76.72	+39.61	0	0
	-76.581	+39.143	-76.72	+39.61	-76.6	+39.4
WILMINGTON	-76.581	+39.143	-76.72	+39.61	-76.6	+39.8
	-76.581	+39.143	-76.72	+39.61	-76.8	+39.4
ELKTON	-76.581	+39.143	-76.72	+39.61	-76.8	+39.6
	-76.581	+39.143	-76.72	+39.61	-76.8	+39.8

After the MOD data input cards are translated by the Dictionary File, all of the input location designations for a data point will be represented by a single LOF code number in the data point record in the Data File. This LOF number consists of the locations of the center points of all the pertinent geographic groupings of the data point. Thus, the various geographic levels may be directly referenced in MOD processing if desired, i.e., the continent, country, province, etc., of any data point can be immediately determined. This method of access does not enhance retrieval, however, since requesting a given country or province, in terms of location, would yield the same results. Direct access to the geographic levels of a data point will be beneficial for some operations which require manipulation and calculation or combination operations.

Each LOF reference number in the LOF code for a location actually consists of longitude and latitude coordinates. Since the lowest (non-synonym) level number provides the most precise geographic location of the data point, it is advantageous to repeat these coordinates in each data point record.

7.3.2 QUERY REQUESTS

With this representation of the geographic location in the Data File, any area can be referenced in terms of the longitude and latitude coordinates of its geographic name. Moreover, the desired boundaries of a map could, theoretically, be expressed in terms of retrieval conditions or output specifications. Thus a map of South American data could be produced by requesting that the output be "South America" (or "Longitude S30 to N15, Latitude W85 to W30") or, more precisely, by specifying that "(LOC) = South America" (or "LON $\geq$ -85 + (LON $\leq$ -30 + (LAT $\geq$ -30 + (LAT $\leq$ + 15)". However expressed, it is obvious that only appropriate data points should be considered in the Retrieval Subsystem.

As previously stated, a query request consists of: retrieval conditions, synthetic or manipulative operations, and output specifications. It

is envisioned that, after implementation of and operational experience with the MOD system, all three of these aspects will be included in a single comprehensive query language, well suited for use by the medical profession. Ultimately, then, the entire query, expressed in that query language, will be interpreted by a preliminary program that will identify and isolate the processing requirements in terms of the Retrieval, Synthesis, and Output Subsystems. But until the MOD query language is developed, each of these subsystems will require its own control-card input -- and the user will have to specify every operation to be accomplished in each subsystem. However, the processing required in the Synthesis and Output Subsystems is often so interrelated that completely separate control cards for these two subsystems would require unnecessary manual effort on the part of the user; furthermore, it would lead to errors of inconsistency. For this reason in the interim system for query requests, it is recommended that the request control cards be limited to two categories: (1) retrieval, and (2) synthesis-and-ouput. The interim system can generate the processing required in both the latter subsystems from a single request entry. Examples of a complete set of synthesis and output control cards will be given after MOD system output usage is discussed.

7.3.3 CALCULATIONS

The manipulative or synthetic operations desirable in the synthesis subsystem are those which can be utilized for both optional calculations and required mapping calculations. These operations can be meaningfully performed upon any single-LOF quantitative MOF (i.e., a MOF whose LOF's are numbers, in contrast to a qualitative MOF, whose LOF's are words). These operations could include the calculation of the total, maximum, minimum, mean, median, and other arithmetic combinations from the numbers contained in any such MOF as found in a group of separate data points.

Some of these calculations require (or can be achieved by) sorting the Retrieved Data File. The selection or rejection of the greatest or least

numerical LOF's in a quantitative MOF is such an operation. Of course the interpretation given to such relative maximums and minimums varies with the requested MOF. For example, in MOF's involving time, the greatest and least values would represent the most recent and the oldest data points, respectively, if the MOF is properly constructed.

Calculations of maximum, minimum, and averages could be accomplished with additional processing in the Retrieval Subsystem for a limited number of MOF's. It is recommended, however, that in the initial development of the MOD system, all statistics be calculated in the Synthesis Subsystem because:

(1) Not all calculations could be performed at retrieval time.

(2) All or some of the calculations may be utilized during the combination portion of the Synthesis Subsystem.

(3) If these calculations are performed for all data points, the basic retrieval operations are extraneous.

(4) Since other calculations may be desired later, there is advantage in (eventually) designing a general-purpose calculating program rather than modifying the (interim) closed Retrieval Subsystem.

(5) Control card formats can be much simplified.

The output of the MOD system can be considered as a summary of certain characteristics of the MOD data -- maps give a pictorial summary, the reports a verbal summary. The desired characteristics are located by the Retrieval Subsystem and then summarized by the Synthesis Subsystem. Each of the synthetic or manipulative operations provides a different type of summary and can be performed on any quantitative MOF for the entire set of retrieved data points. Moreover, these operations can also be performed for any well-defined homogeneous subset of the retrieved data. Consider these examples: (1) the average number of cases of a specific disease could be determined with respect to all of the data points, each of which had all the other

desired characteristics; (2) the average prevalence or incidence of a particular disease for each given year could also be ascertained from a group of data points having the necessary elements, and this average could either be based upon all the data points or upon only those from times when there were no epidemics occurring.

Illustration (2), above, is an example of performing a calculation on a single-LOF quantitative MOF, say A, for each LOF of another MOF, say B. For simplicity of expression let us represent this type of operation by f(A):(B), where f is any defined calculation. If the calculation is to be performed with respect to the entire file, let us denote this operation by f(A):(#) for compatibility. B need not be numeric for the operation to be meaningful since the calculation is performed for each different LOF in B. The usual data processing technique employed to accomplish f(A):(B) is to sort the entire (Retrieved Data) File by B and then to calculate f(A) for all A's which have the same B. Thus, in effect, a separate calculation is performed every time B changes. In this calculation, it is assumed that there is only one MOF "A" and MOF "B" in each data point record, hence f(A):(B) is an inter-record calculation.

It may be desirable to perform calculations on several single-LOF quantitative MOF's within each data point record. This is an intra-record calculation and can be denoted by f(A,B,...,X).

In the initial Synthesis Subsystem it is recommended that a general-purpose CALCULATE program be written to compute only the following calculations:

Calculation	Suggested f-designation
TOTAL	TOT
MAXIMUM	MAX
N-GREATEST	MAXN
MINIMUM	MIN
N-LEAST	MINN
MEAN	MEAN
MEDIAN	MED
STANDARD DEVIATION	SD
ADD	+
SUBTRACT	-
MULTIPLY	*
DIVIDE	/

The CALCULATE program would contain each defined calculation as a subroutine, thus meaningful combinations of these calculations would be relatively simple to perform. However, an order-of-operation or parenthetical-grouping standard or convention would have to be established to make these combinations well-defined.

Since the results of these calculations must be transmitted within the system, we will now consider their internal representation within the system.

Calculations which are performed with respect to the entire retrieved data file will be contained in a generated last record of the data file.

Intra-record calculations will be stored in a new MOF of the data record and given a MOF designation equivalent to their (calculation) f-designation, also indicating what MOF's were involved in their calculation. The resultant calculation will also be stored in main memory for possible utilization as an operand in another calculation.

Inter-record calculations can be constructed in either of the following ways:

(1) All of the original data point records plus summary records.

Each summary record would contain f(A) as the LOF for the MOF A, and would immediately follow the group of data point records to which it pertained.

(2) Summary records only.

In these records, any subset of the summarized MOF's, i.e., all MOF's which were constant during the calculation, plus f(A), could be written for each resultant calculation. This second type would require an additional specification to f(A):B, which would indicate those MOF's to be included in the summary record.

7.3.4 SORTING

The specification of the MOF B is sufficient to accomplish f(A):(B). Actually, however, B may be the most minor MOF of several MOF's into whose sequence the Data File was sorted. For the present, it is suggested that all of the MOF's required for the proper sorting sequence be explicitly stated on a sort card. In the ultimate MOD query language these sorts can be automatically generated from the calculation specifications. In addition to their utilization for calculations, sorts will be effected in order to achieve a desired order in output reports. Unlike map output, the significance of printed alphanumeric reports can be greatly enhanced by a provision to vary the order in which the desired data points are listed. The sorting does not require an additional computer run if the LOF's (or MOF's) are to be printed in the form of their textual descriptions. But in this event, the Data File must be matched against the Dictionary File to obtain the proper descriptions. This process can best be accomplished if the data records are decomposed into MOF/LOF records and sorted, hence, the resultant file, containing the textual descriptions, would have to be re-sorted. Only those MOF's which are to be listed need be decomposed and sorted. The MOF's to be decomposed, and their desired output order, can both be obtained

from the output specifications without additional synthesis control cards in the interim MOD system.

The nature of the MOD Data File requires that two features of its structure be given special consideration in sorting. First, the variable locations of each MOF in the file are not suited to standard sort programs. This difficulty can easily be rectified by reformatting the data point records so that all the MOF's to be sorted are placed in fixed locations as they enter the first pass of the sort program. Secondly, a procedure must be established for sorting multi-LOF qualitative MOF's. Although quantitative MOF's can certainly be defined so that they are single-LOF, it is often desirable to have multi-LOF qualitative MOF's. As the system has been designed, all of the LOF's within a single MOF will have equal significance, thus only the following two methods of processing multi-LOF MOF's are feasible for sorting MOD data:

(1) Sort on the first LOF for each MOF but retain the other LOF's. The processing in the MOD Storage and Retrieval Subsystems will cause the several LOF's to be sequenced by increasing LOF code number within each MOF.

(2) Create an entire new data point record for each of the several LOF's within a MOF. All of these records would contain the same data plus a generated flag.

Either of these methods could be specified for each (qualitative) MOF which is to be sorted. For the present, the user would indicate the better technique after considering the structure of the MOF and its contained LOF's, the calculations required, and the type of printed or mapped output desired. Both techniques could be used in the same sort program for different MOF's. The general form of sort control card to sort the file by MOF's A,B,C, and D, respectively, could be SORT (A,B,C,D), where A,B,C, and D are the MOF

designations. In the second technique for sorting multi-LOF MOF's, those MOF designations could be prefaced by an asterisk "*". If no choice were indicated, the first technique would automatically be employed. Often the sort control will be supplied from the output specifications, and the same conventions can apply in a print control statement. If a single-LOF quantitative MOF is chosen for inappropriate processing, the sort statement will be flagged and the option will be ignored for that MOF. Later embellishments to the MOD system could assign the better method for each MOF automatically, and could include options to reformat the entire data point record prior to its being sorted.

7.3.5 COMBINATIONS

Thus far we have considered summarizing operations which are optional for either printed reports or maps. Certain summarizing operations will always be required in order to produce a meaningful map from the remaining retrieved data points, however. These summarizing operations include the following:

(1) Making the data point values consistent in form so that they may be meaningfully compared and calculations performed on them.

(2) Combining all data points which possess identical LOC's (locations).

(3) Combining, then, all such points which will be mapped at the same grid point.

Many possible ways of combining data points can be envisioned; most of them reduce, essentially, to taking some sort of average LOC (location), and coupling it with an average of the VAL's (values) of the data points.

The MOD Data File consists of data points extracted from medical papers in which the degree of geographic significance will vary and in which the geographic areas will be inexact. There may even exist data points which reflect contradictory data, i.e., all their independent LOF's/MOF's and LOC

are identical but the values (VAL) of the points are different. Such contradictory points could be purged during MOD Storage Subsystem processing or allowed to remain in the Data File and then "combined" in accordance with the established combination techniques. It is anticipated that many of these "extra" data points will be eliminated from consideration for a specific map output by the retrieval requests and processing. However, any points for the same location which remain after such processing must be combined prior to mapping.

The geographic points which must ultimately be combined to produce a mappable point depend upon both the lowest level of geographic unit to be considered and the desired grid size. For a given area the user may wish to vary grid size since, as previously demonstrated, different grid sizes can produce dissimilar maps. The size of the area to be mapped can also influence the desired grid size For example, maps of the world, or of a country, or of a state would probably be automatically assigned grid-mesh sizes of 1°, $1/2^{o}$, $1/10^{o}$, respectively. Any grid size larger than the smallest grid size contained in the location part of the Dictionary File may be constructed by combining all data points at the closest new grid point location.

All of these necessary combinations of data points can be achieved by use of the previously discussed calculation program (CALCULATE) of the Synthesis Subsystem. When this program is used to effect final synthesis for mapping (during initial MOD implementation), we suggest that no new functions be defined for this purpose. Later, a new algorithm may be developed which will optimize this operation, in which case that algorithm can be added to the calculation capabilities of the entire Synthesis Subsystem, and can be made the standard procedure unless another method is explicitly specified by the user.

In our discussion of the MOD gazetteer and grid functions, it was noted that once the LOC (location) had been established for a data point record,

all defined geographic levels of this location could be referenced if desired. These levels appear to be a suitable criterion for the geographic combination of data points. For example, a map displaying the total number of cases of a specific disease per province could be produced by summarizing the data point values (VAL) at the province (PRO) level. This calculation could be performed by the synthesis statement TOT(VAL) : (PRO) after the retrieved data points had been sorted by location. A map of the same data by county (CTY), and with the grid size decreased to 0.5°, could be achieved by TOT(VAL : (CTY.5). In this calculation any data point whose location did not include a county specification would, of course, be omitted from the total.

In the MOD System, the mappable value (VAL) of a disease data point may be represented in terms of both absolute numbers and rates or percentages, and the final combination of data points for mapping will often include the requirement that these percentages be combined. Although the combination of percentages is less well-defined than that for absolute numbers, various types of such combinations can be calculated if the values and sample sizes of the data points to be combined are known. For example, 50% and 10% can be combined to yield a value of 10.784% if the respective sampling were known to have yielded 1 out of 2 and 10 out of 100 cases positive. However, the same percentages could also have been combined to values of 30% or 60% depending upon the combination technique employed and the sampling situation involved. In the MOD system, the data analyst will probably be the one (initially) to specify the best method of combining pertinent percentages. For example, the combination method could be specified by TOT(VAL)/TOT(SAM) : (PRO) if VAL and SAM were the MOF designations for Value and Sample Size respectively, and a grouping by province were desired.

Often, the sample size for the disease measure cannot be determined from the data included in the published report. For this reason, MOF's for the largest and for the smallest sample sizes, (LSZ) and (SSZ) are

contemplated. In this event the data analyst must also specify the method of calculating the sample size to be used. If in the example given above, an average of the largest and smallest sample sizes were desired, the data analyst would specify TOT(VAL) / TOT(MEAN(LSZ, SSZ)) : (PRO). Of course if the sample size were actually known, both MOF's would have to contain the same numeric value if this calculation were to be well defined for all data point records.

7.3.6 ENHANCEMENT

For the production of maps, the final combination operation must reduce the MOD data to longitude-latitude coordinates and values. This collection of points may then be enhanced by the addition of other points as the final processing step in the Synthesis Subsystem. The techniques by which contour and shaded maps will probably be drawn require that those points in non-applicable areas which fall within the longitude and latitude range of the map to be produced be added in with the set of previously processed (retrieved and combined) data points. These points can be added by obtaining the non-applicable locations from the dictionary (grid) for the area to be mapped. In addition, if a shading map is to be drawn, those grid points which fall within the area under consideration, but which are not contained in the retrieved data, must be added in order that the geographic extent of the area be properly represented. Each of these points must be given the same value as the appropriate retrieved data point, and their locations can be obtained by matching the retrieved data points against the dictionary (grid). (The same technique could be applied to contour maps, but it is more common to use only the center points of each geographic area for such maps.)

The inclusion of certain non-applicable points could be extended to all types of maps to provide a graphic representation of the area boundaries. At the present time, however, this process is unnecessary since these boundaries would be evident on the base map (upon which the MOD distribution map is to be overlaid).

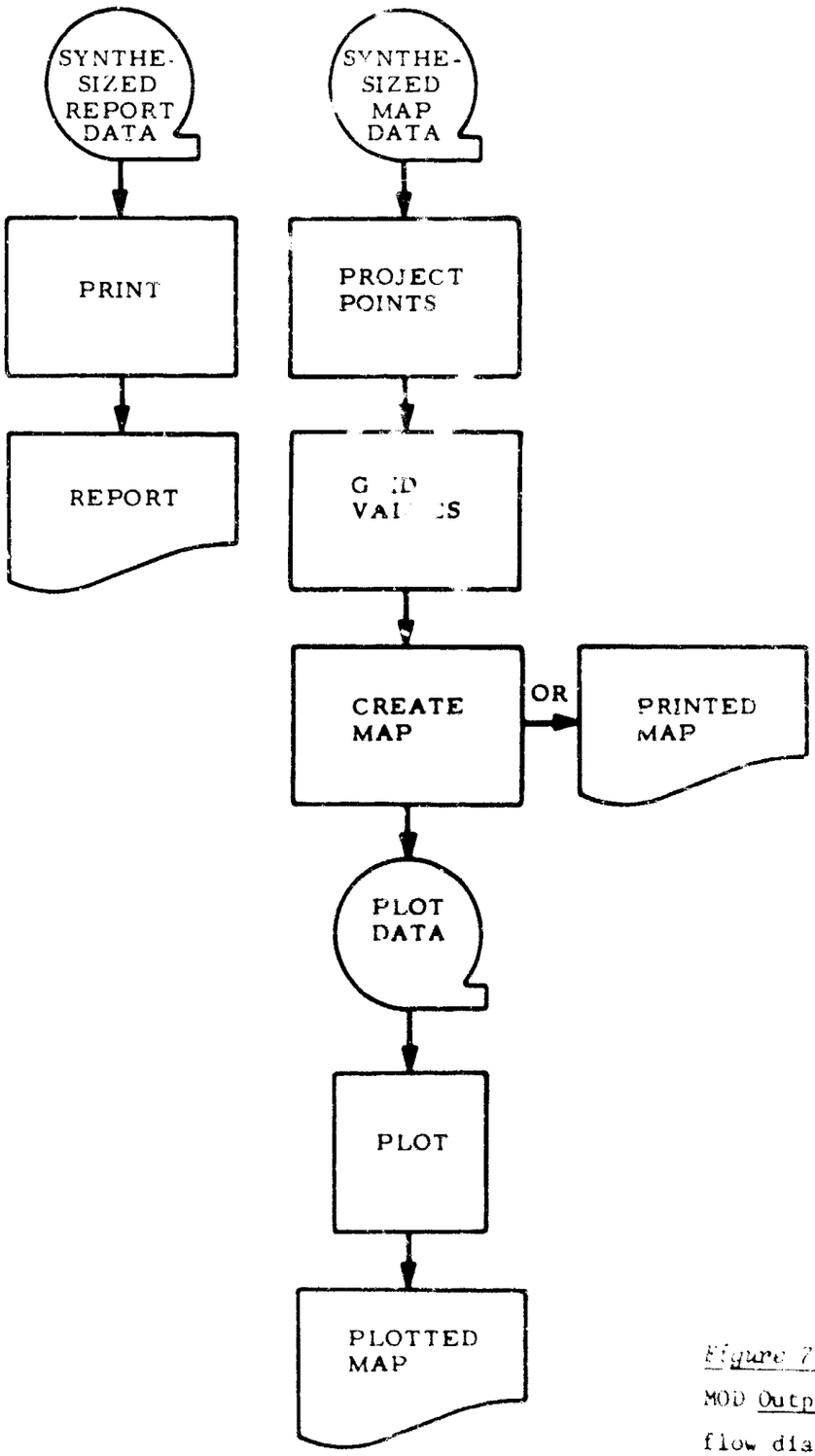


Figure 7-6

MOD Output Subsystem flow diagram.

7.4 OUTPUT SUBSYSTEM

The objective of the MOD System is to display medical data in terms of geographic distribution. In order for the stored, retrieved, and synthesized medical data to have significance for the user, the resultant data must be meaningfully displayed. The Output Subsystem (see Fig. 7.6) provides this display in the form of maps and reports. The Storage, Retrieval, and Synthesis Subsystems will have produced internal records which contain the required data, thus the Output Subsystem is the least complicated both conceptually and structurally. (Actually, most of the output considerations had to be formulated initially in order to design the other three subsystems.)

7.4.1 REPORTS

Reports will provide an alphanumeric or verbal listing of selected MOD data. One program should suffice to accomplish all printing requirements. Printouts will normally consist of several pages. Each page will include a brief heading and the page number, and, at least the first page, will also contain the entire query request. Usually, the textual description of each desired LOF will be listed, but provision should also be made to list only the LOF code number. This latter type of report would be particularly useful in the early stages of implementing the MOD system since it would eliminate the additional processing required to convert the data file LOF codes into LOF names. In addition, the user should be able to request that MOF's be listed by their textual description or by their MOF code designation. As previously mentioned the conversion of MOF's or LOF's to their textual descriptions must be accomplished by the Synthesis Subsystem to assure proper continuity within the system. In either event, any question marks associated with particular LOF's would be printed.

For additional flexibility, the MOD output reports can be furnished in either of the following forms:

(1) Free form -- An entire data point or any portion of it can be listed in free form across the printed page. If the user specifies the MOF's to be listed, he also indicates the order in which they are to be printed from left to right. The first designated MOF will protrude to the left, for each data point. If no MOF's are specified, all MOF's will be listed, and they will appear in the same order as in the data point record with the data point number first and the narrative last. A MOF description will immediately precede its contained LOF's. If MOF code designations are used, the MOF and LOF entries will be listed in free form across the page. However, if the textual descriptions of the MOF are desired, each line will probably contain only one MOF and all of its LOF's.

(2) Fixed form -- A tabular listing of portions of a data point is printed on each line of the report. For this type of report the user must not only specify the MOF's to be listed but must also indicate the maximum number of characters which he desires to be allotted for each MOF. The MOF's will be spaced across the page automatically, therefore the total maximum number of characters, plus at least one space between each MOF, must be no greater than the total number of print positions across the page (usually 132). The right-most characters of a LOF will be omitted if its textual description exceeds the allotted number of characters. The first line of each page (after the heading) will consist of the MOF titles for each column. Again, these can be either MOF code designations or textual descriptions. The latter descriptions will also be truncated if there is insufficient allotment for character length.

In fixed-form reports only the first LOF in a group of LOF's all belonging to the same MOF will be listed; all such LOF's will be printed in a free form listing. For this reason, if the report is to be printed in fixed form, ordinarily the user should elect to create multiple records in the SORT program of the Synthesis Subsystem.

Calculations that were performed in the Synthesis Subsystem will also be listed in output reports. But in the initial system, the data analyst must insure that all sorting and translation is accomplished prior to inter-record calculations in order for the summary records to appear after the appropriate group of data points. Inter-record calculations will be printed as separate lines in either fixed-form or free-form reports. Intra-record calculations will appear with the name of the calculation as its MOF title within the record print-out for either type report, if the calculation is designated in the output specifications. Calculations for the entire set of data will be listed last.

7.4.2 MAPS

Maps (and the very similar block diagrams) will provide a pictorial or graphic display of the selected MOD data. In the modular development of the MOD System, experimentation with existing mapping techniques and selected subsets of data did not proceed to the point of yielding final or definite data-processing solutions to <u>all</u> mapping aspects and problems. However, it is certain that the production of a finished map from the synthesized data will (ordinarily) require three steps:

(1) Project the enhanced data points in accordance with the projection specified (generally the same as that of a base or environmental map with which the MOD disease map will be compared).

(2) Grid the projected locations of these points for contour and shaded maps.

(3) Produce a map with a plotter.

7.4.2.1 Projection After the final synthesis and enhancement of the data points, these records will contain two elements: the location (longitude-latitude) of each point and the value (of that point) to be mapped. But these longitude and latitude coordinates cannot be directly transcribed onto a meaningful map since these coordinates are actually two-dimensional locations on a three-dimensional spheroid. Maps are conventionally constructed from the projections of these coordinates, except for maps of very small areas. Among the most commonly encountered map projections are the Mercator, Miller cylindrical and Goode's homolosine projections. Most of these projections will be useful for the MOD system since each displays a better pictorial characterization of earth areas and distances under different circumstances. Moreover, since the MOD-produced maps are to be overlaid onto existing environmental maps, various projections of the MOD data will be required to make this data correspond spatially or areally to the environmental data. Formulae by which the longitude-latitude coordinate values can be transformed into X-Y coordinates for any of these map projections are readily available -- in fact, there are existing computer programs which will perform most of these transformations.

7.4.2.2 Gridding After the MOD data point locations have been transformed by the appropriate projection, existing computer mapping techniques require that the MOD data be gridded to produce either a contour or shaded map. Gridding consists of constructing an array of new points (the vertices of regular polygons) from the existing points. These polygons are most often squares, rectangles, triangles, or hexagons, and are called grid boxes. These grid boxes usually are constructed to have equal areas, although a variable grid size is occasionally used. Each point on the new grid is assigned a value by interpolating between the values of those data points relatively near the new grid point. In some techniques the gridded area is smaller than the original area, in others it is slightly larger. Each interpolated value can be calculated by methods which range from a consideration of only two of the original values to those which include

every original value, and methods which involve simple linear interpolation to complex non-linear interpolation. Use of the nearest five to eight original data points appears to be optimal.

The grid-point values obtained by these techniques give values to the closest integer suitable for the production of contour maps. However, these methods would have to be modified for producing shading maps, providing for each grid-point value to be the same as the original value for that area. It may be desirable to grid non-data-valued locations in this latter manner for both contour and shaded maps.

Grid criteria can be established so that dot, shaded, or contour maps could be produced on a line printer. Each grid point would have to be one of the print locations on a page, and the scale could not be varied from 10 x 6 or 10 x 8 points per square inch. Under these conditions contour maps would have to be represented by groups of the same print character rather than lines, in which case shading would consist of discrete characters rather than continuous symbols.

7.4.2.3 Production of Maps If MOD maps were to be produced on a printer, the processing required to produce maps from the projected and gridded MOD data would be relatively simple. However, it is envisioned that MOD maps most commonly,would be drawn with an automatic (digital) plotter. The actual plotting operation is almost always accomplished off-line, i.e., a magnetic tape is created during the system processing, and, subsequently, this tape is used as input to the plotter device. The magnetic tape consists solely of a series of X-Y plotter coordinate points and an indication of whether the plotter pen is up or down between these points.

The conversion of the grid coordinates to plotter coordinates determines the scale of the resultant map. The plotter coordinates are expressed in X-Y values with accuracy from 1/100 to 1/500 of an inch. One inch on the plotted map can be equivalent to a varying number of miles in different

areas of the earth, depending upon the type of projection. The scale of miles, in terms of inch equivalent, is obtained from standard reference lines in each projection. The desired scale must be specified by the user and should conform to the scale of the existing environmental map with which the disease map will be compared. The conversion of the grid points to any scale is always a linear transformation. Obviously, the maximum dimensions of the plotter page must not be exceeded. For drum-type plotters, one dimension can be indefinitely long, but both dimensions are restricted if a flat-bed type plotter is used.

The plotter instructions to produce any desired legends, numbering, and register marks is also represented on the plot tape by a series of X-Y coordinates. For plotter efficiency it is recommended that the identification and the legend drawn with the map be brief, and that lengthy groups of characters, such as the entire query request, be listed on a printer.

A computer program is necessary to create the plotter tape from the gridded (or projected) data points. For contour maps alone, many such programs already exist, but each was designed for a specific application unlike that of the MOD project, and we do not yet know which program would be the most generally suitable for MOD purposes. These various programs produce quite dissimilar maps with the same data. Although some existing programs have worked well with certain MOD data, it is not fully apparent yet whether it will be more desirable (ultimately) to modify an existing program or to design and implement an entirely new one.

Different programs will probably be required to produce various types of MOD maps. Processing methods and other considerations for each type of map are as follows:

(1) Dot-type maps: In dot maps, the value for each point can be appended to the point's location. Zero values can be indicated to contrast with unknown values. Dot maps may

also be used to illustrate absence or presence without any indication of numerical values or rates. Each point and its value are drawn directly from the projected (scaled) locations.

(2) Shading-type maps: For shading maps indication of the interval values and the symbols which are to distinguish the value levels representing each such range must be supplied by the user. Since these indications will be punched onto cards, special provision must be made in specifying non-standard computer characters for the shading symbols. A common plotter practice is to predefine these non-standard symbols by numbers or short words. These numbers or words are then used in the symbol designations. Later, the symbols can be assigned automatically in a standard sequence of increasing density. Non-applicable areas could appropriately be assigned a special symbol. The map would be drawn from the (scaled) grid points. Each grid box could be shaded individually. Alternatively, adjacent boxes possessing the same shading value could be shaded at one time. (This would require some additional processing, but would substantially speed the plotting operation.)

(3) Contour-type maps: The desired contour intervals must be provided by the user. Each interval could have an equal increment or each increment could be explicitly requested. The usual contour technique is to draw all the appropriate contour lines within each grid box, one at a time, and then proceed to the next grid box. Provision must be made to end the contours at user-specified, non-applicable locations obtained from the Dictionary File. The values of the contour lines could be indicated on the map (Shading and contour intervals

for certain data might be too small to be read. Such a situation could be determined prior to mapping and remedied either by terminating, with an appropriate message to the user, or by automatic selection of a more suitable type of map.)

(4) Combination-type maps: Some combinations of these types can often produce more meaningful maps than can a single type. A contour map which also indicates the original data points would be of value where the contours were a better representation of reality. A contour map in which the areas between contour lines were shaded would graphically relate similar values and distinguish peaks and valleys. It is theoretically possible to combine meaningfully shading and dot maps, but there are technical limitations since their representation by the plotter would often be unreadable.

7.4.3 MULTIPLE OUTPUT

A map presents a pictorial description of the retrieved data, but this portrayal is limited to location and value. It would often be desirable to augment a map with a verbal description of some pertinent MOF's associated with each data point. A report accompanying the map could describe the data points in terms of the summarized points, its component points, or both -- or the narrative accompanying its component points.

8 Output usage

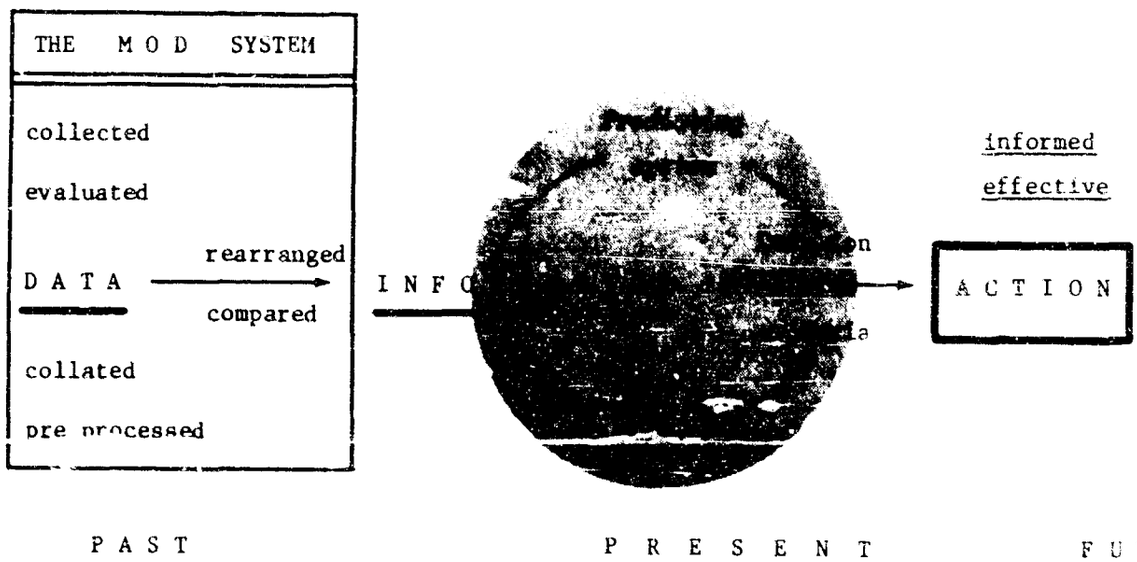
ABSTRACT - This section discusses operational procedures and considers how the MOD system can be most effectively used. In the "Notes to user", inherent limitations of the map form as a means of presenting information are discussed, also restrictions imposed by the data base. Potential applications of the MOD system are considered, and several examples are given.

"The interpretation of knowledge must take ignorance into account."

Professor Levy

8.0 GENERAL CONSIDERATIONS

The (implemented) MOD system, including its data bank, is simply the means to an end. It must be used effectively to give insight into disease/environmental situations, helping the user to arrive at informed decisions which will lead to appropriate action.



The true purpose of knowledge resides in the consequences of directed action.

John Dewey

Continuing experience with the operational system will be necessary to reveal all of the ways in which the MOD system can be used effectively. Obviously, the details of such usage cannot be given now, but as a guide to our development of the system, a basic pattern of output usage was formulated.

8.1 OPERATIONAL PROCEDURES

The MOD system is unique and sufficiently different from other systems to require detailed instruction. This instruction will be provided by a "user's manual" which will explain the language and the detailed procedures for operation. The major steps are as follows:

(1) User conceives of an idea or hypothesis that he wants to test with the MOD system.

(2) User writes out a rough-draft preliminary query, including retrieval conditions consisting of: disease/environmental factors, and geographic locations/areas, and synthesis, and output specifications for the kind of map desired.

(3) Data analyst, in conjunction with user and/or data consultant, rephrases query in terms/format acceptable to MOD system.

(4) Query is keypunched.

(5) Query is batched with others and fed into system. (errors are returned and corrected, then re-entered by using procedures previously outlined in steps 3 or 4.)

(6) MOD system retrieves data points from Data File, manipulates them, and produces maps (sometimes accompanied by supplemental reports), each showing the distribution of the areal variations of one disease/environmental factor.

(7) User takes the maps and compares them, ordinarily by overlaying them on each other and on (published/drawn) base maps (taken from the MOD map library) to determine pattern fit, including, perhaps, variations in pattern related to year, season, etc., etc.

(8) User observes new interrelationships (<u>not</u> new data) and gains new perspectives and increased understanding of the disease/environmental situation (e.g., salient disease-environmental relationships discovered/confirmed/disproved and/or pertinent modifications that need to be made in data collection/data files and/or better ways to phrase the old query or to formulate a new query in order to generate additional information).

continued next page

(9) User draws conclusion, makes (informed) decision(s) and initiates whatever action(s) is deemed desirable/necessary.

Figure 8 - 1 illustrates, in schematic fashion, the various steps that are followed in the MOD system: collecting <u>data</u>, preprocessing it for computer input, manipulating it in response to query, and outputting it as <u>information</u> in accordance with users specifications.

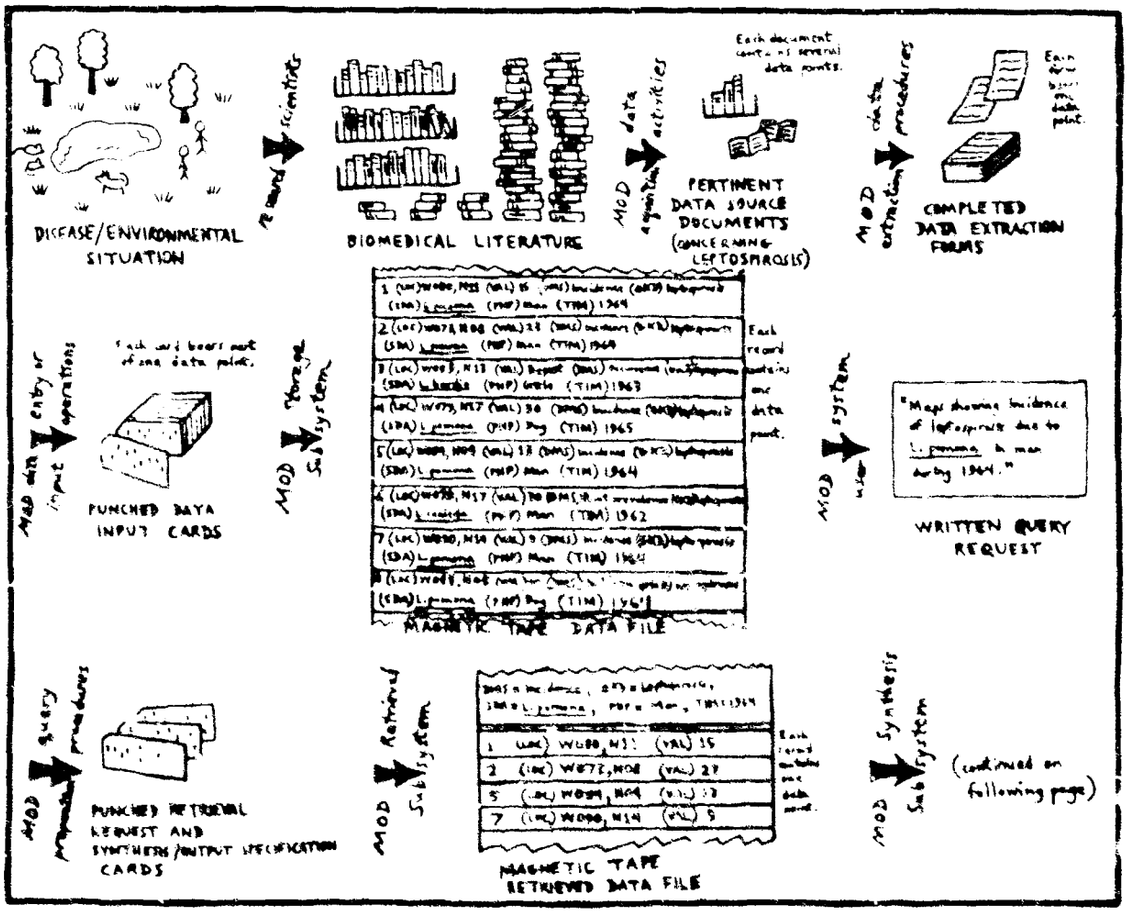


Figure 8-1 Overall pattern of MOD system usage.

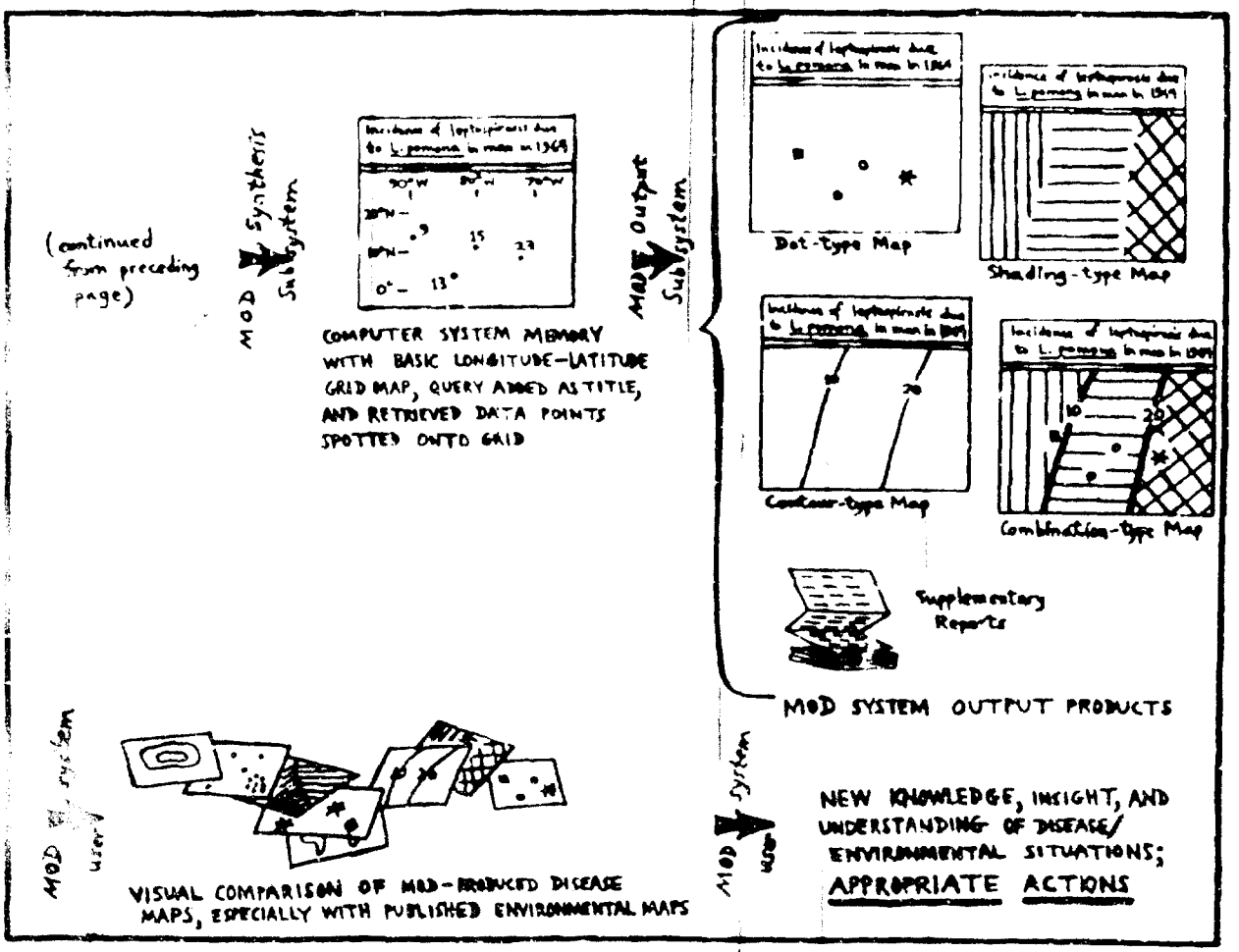


Figure 8-1 continued.

8.2 NOTES TO USER

In the MOD computerized system it is the user, not the system, who makes correlations between the raw data and output map, evaluating the various factors which make the map look as it does. The computer system will not perform analysis of the maps produced nor will it make judgments; it will merely manipulate (according to rigidly defined algorithms) extracted, formatted data (from that pool of data which was previously put into the system) and output these manipulated data in the form of maps or other reports, in the manner specified.

As has been discussed before, it is the mandatory responsibility of the user to understand maps and their use, in general, before attempting to interpret specific maps produced by any system. The potential user of the MOD system will require considerable orientation and training in three areas: logic (to pose the query); cartography (to understand what a map is, what it can do, etc.); and the biomedical disciplines (to understand the limitations of data, including what kinds can and cannot be manipulated and mapped). The effective use of maps involves competence on the part of both the compiler and the reader with respect to three fundamental factors: an understanding of map scales, how to determine position, and how to present the data in a form that can be readily assimilated.

For example, various situations may all result in similar-appearing blank areas on a disease-distribution map. The MOD system user must be aware of several possible causes of such blank areas if he is to interpret the map correctly. Some of these possible causes are:

(1) The disease was looked for but found to be absent. (Ideally, this is what all blank areas should indicate, but this ideal is a very long way from fulfillment.)

(2) The disease never been looked for, or diagnosed, or reported from that region (but may be present there).

(3) The disease is present but has been incorrectly diagnosed and reported as something else.

(The region mapped is uninhabited.

5. Output Usage

Before making his query the potential user of the MOD system should be required to check the MOD Map Library catalog to see whether or not the map he wants has already been requested and ouput for a previous worker.

We suspect that, in the study of disease and environmental situations, assuming full operation of the MOD system, most of the disease maps will be computer-produced, while most of the environmental maps will either be found already published in suitable form or will be manually-produced from data presented in books, atlases or major reports.

When the user (usually a professional biomedical person) compares the maps, he will do so visually on a trial and error, i.e., subjective basis. For example, he will look at and compare a map showing the distribution of leptospirosis with another map showing the distribution of rainfall in the same area and conclude, perhaps: "leptospirosis is related to rainfall". His basic operating assumption is that a similarity of distribution patterns on maps implies some relationship among the factors mapped.

Maps produced by the computer can be output on transparent material which can then be combined (matching geographic points) with other factor maps. The practical limit to the number of overlays is probably quite low, for the whole purpose of this type of data processing is to simplify the situation being considered so that relationships are clarified. If the patterns exhibited by the disease and environmental factors are similar (i.e., they match), some relationship can be assumed between the disease and environmental factors. However, only further study can determine the nature of that relationship -- whether it is causal or, merely associative. Figure 8-2 illustrates the various types of such relationships.

Other ways by which existing maps could be compared with data contained in the MOD system data files involve manipulating the existing map: The data it contains could be digitized and input to the computer files.

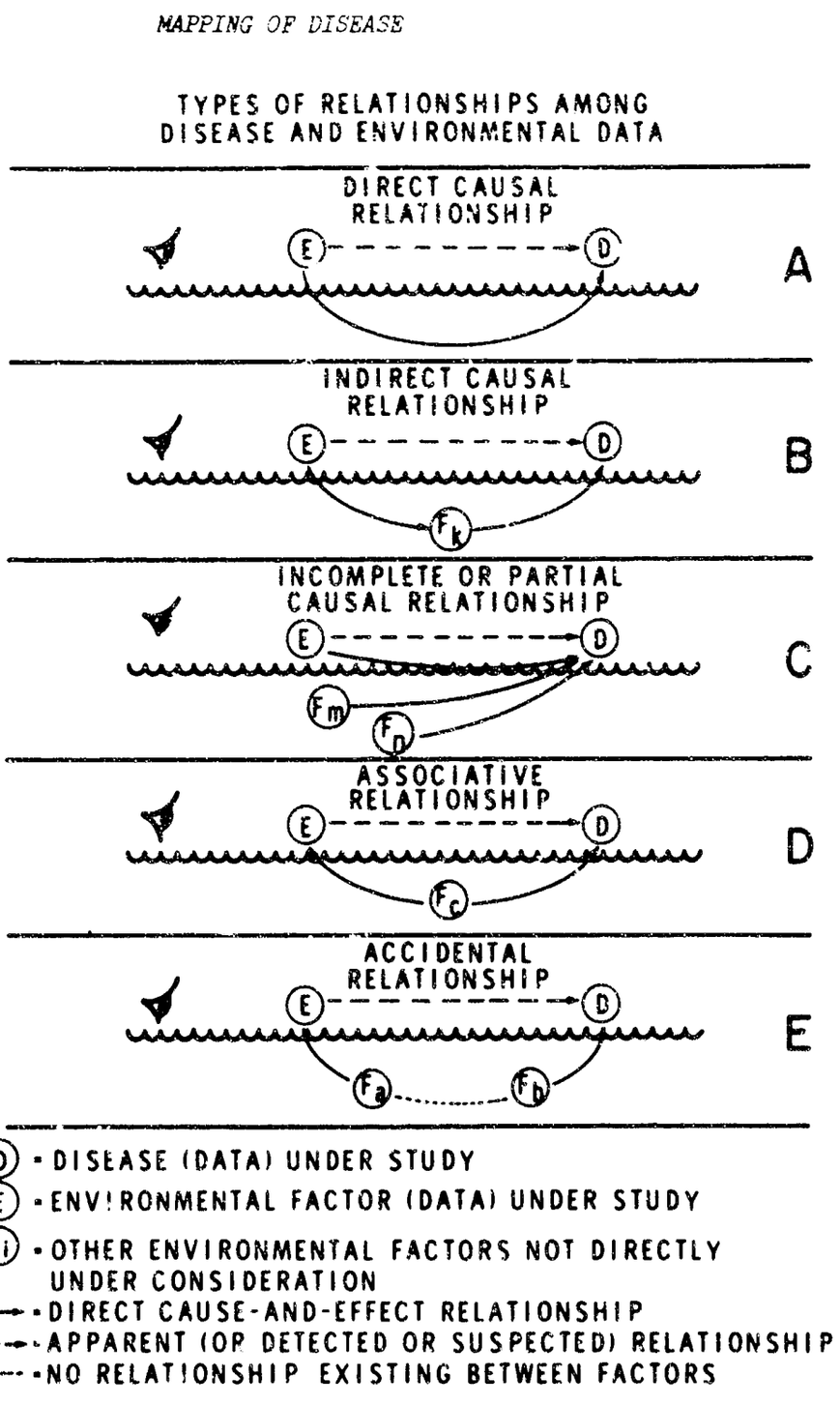


Figure 8-2 Types of relationships among disease and environmental data. The eye of the observer is evident at the left. Those connections above the surface are readily seen; those below the surface are not.

The map could be redrawn manually to a different scale and photographically reduced or expanded to make a new hard copy of the appropriate scale. A relatively simple and inexpensive way of comparing maps of the same projection, but different scale, is to make (photographic) transparencies and project each of these on the same screen simultaneously, using separate slide projectors, adjusting projector distances so that the maps superimpose. Alternatively, the MOD map could be manipulated, or the data existing in the MOD data files could be mapped on a new projection, scale or other basis to fit the base map.

Overlaying and visual pattern comparing is a very powerful process because it permits human detection of relationships which are so complex that standard mathematical methods would be unable to detect them. Used in conjunction with a computer, the process of map preparation is greatly improved, as the user can get an up-to-date map, i.e., distribution pattern (as far as recorded data is concerned) within a few hours. The user might want to "clean up" manually parts of a computer-produced map, but this is relatively simple as compared to preparing the whole map.

8.3 POTENTIAL APPLICATIONS

It is appropriate once again to emphasize that the major objective of the MOD project is to develop a system whereby narrative and tabular data can be collected and preprocessed (formulated) so that they are suitable for subsequent computer processing and output in the form of distribution maps, graphs, tables, and narrative. Although the self-imposed limitations described previously (input data has mainly concerned the ecology of schistosomiasis and leptospirosis) narrow the limits of specific output considered in this report, they do not narrow the potential limits of the system. The system has been designed to meet certain needs for information dealing with infectious disease, however, the same system could be used, with little modification, to analyze the ecologic factors which influence efficient stockpiling of corn or aluminium, or the ecologic factors which influence

efficient forest preservation or development of recreational facilities, or the ecologic factors which influence efficient development and location of community blood banks or Medicare treatment centers, etc. etc.*

➡ The value of a computer system that allows rapid presentation of current or historic disease/environmental information in tabular, graphic, or map form is so obvious that it requires no elaboration. But there are other, less obvious uses of the MOD system.

➡ For particular diseases, in relation to particular geographic areas, the MOD system can provide a very valuable research tool since it makes possible the rapid presentation, in a vivid way, of relationships among disease, per se, man, and his environment so that the ecology of disease becomes more clearly evident, and causally related factors more readily apparent.

➡ Through correlation of many causal factors in relation to the current situation and the recent past, the MOD system would enable the user to determine trends and, in this way, to get a reasonable perspective of what the future might be.

➡ The MOD system provides much insight into the minimal requirements of disease data in order that these data can be computer processed. Thus the system becomes helpful in preparing data extraction forms for any disease/environmental situation. In this way, the MOD system can give excellent support for anyone wishing to carry out a <u>prospective study</u>. It can also be of considerable value in suggesting ways to evaluate data already on hand, and in determining the feasibility of a <u>retrospective study</u>.

---

* Very recently AID (Agency for International Development) has expressed great interest in the MOD system as a means of identifying, and characterizing, and locating (on maps) those disease-environmental situations which would probably interfere seriously with proposed schemes for economic development of several Latin American Countries.

8. Output Usage

As an important by-product, the mere <u>existence</u> of a computer system which can manipulate disease environmental data to yield valuable information will provide an important stimulus to get more <u>and better</u> data -- data that are more nearly complete as well as more accurate. Furthermore, the MOD system, by pointing out "bare areas" in the data pool, will direct attention where it is most needed.

There are two principal ways in which the MOD system could be used to investigate causal (ecologic) relationships in infectious diseases. First, one could take a set of variables, the values of which were actually recorded in relation to a particular disease situation, then determine the relationships which <u>did</u> exist. Alternatively, one could select a number of variables thought to be important, then alter these (systematically) to see if the information ouput was consistent with what might reasonably be expected, i.e., whether or not the results made medical sense. Obviously, both of these approaches have their place:

(1) To take what <u>did</u> happen and try to determine why (in the sense of identifying dependent variables).

(2) To develop a hypothetical situation and attempt to predict what <u>might</u> happen under those conditions.

Many specific kinds of questions could be put to the MOD system, for example:

- Given particular environmental changes, what changes in incidence/character of a specific disease are apt to occur?
- Given the past history and broad trends of a particular disease/environmental situation, what is the likelihood that major variations in incidence (i.e., epidemics) will occur within the foreseeable future?
- Given particular changes in a disease situation, what specific environmental factors might have caused or influenced these changes?

continued next page

- Given several environmental factors, which one(s) are most likely to influence distribution of various animals which may act as intermediate hosts or reservoirs of disease or -- on the other hand -- which may yield economically valuable products such as pearls or furs, or food?
- Given several different diseases, what interrelationship, if any, exists among them? For example, among protein malnutrition, iron deficiency, tuberculosis, and hook worm infection or between influenza and (subsequent) pulmonary emphysema, etc.

Obviously, the output of the MOD system is "information," information directed primarily toward helping bio-medical scientists:

(1) Appreciate more fully quantitative aspects of disease/environmental data in relation to place and time.

(2) Identify the multiple causal factors of a given disease and their interrelationships.

(3) Determine interrelationships, if any, among several different diseases or conditions occurring together.

(4) Evaluate the impact of the disease upon socio-economic aspects of the area, military operations, etc., etc.

(5) Anticipate the effects of altered ecology on incidence and manifestations of disease.

(6) Predict variations in incidence and changes in character of disease that are likely to occur in the foreseeable future (on the basis of past history and trend analysis).

3.4 EXAMPLES

In developing the MOD system, operation was simulated using real data, data that reflected realistic situations. Many of these operations were

limited simply to mapping incidence of a specific disease, but other, more complex situations have also been explored, as illustrated by the following examples.

Data concerning the distribution of Burkitt's tumor has been redrawn into the form in which it would be output by the MOD system (Fig.8-3). One base map of Africa is shown (Fig. 8-3A), on which two other maps may be superimposed. One of these maps (Fig. 8-3C) shows the occurrence of Burkitt's tumor -- the other (Fig. 8-3B) shows those regions in Africa where, simultaneously, the altitude is under 5,000 feet, the seasonal mean temperature exceeds 60°F, and the total annual rainfall exceeds 20 inches. When these maps are overlaid they give the appearance shown in Fig. 8-3D.

Data dealing with the distribution of goiter and the iodine content of water in the United States provide a second illustration of these techniques (Fig. 8-4).

For a third example of MOD system usage, we returned to the standard set of schistosomiasis data used previously in testing the various computer-mapping programs. Again, we emphasize that this example is offered only to illustrate technical aspects. With the restrictions imposed by the limited data being used, one must not draw firm conclusions about the disease-environmental relationships.

Assuming that a user is interested in the relationships among infection rate of schistosomiasis, rainfall, and temperature in eastern Brazil. He can ask for separate maps, each showing one of these factors (Fig. 8-5A,B,C), then overlay them (Fig. 8-5D) to compare their distribution patterns. From this it appears that July normal temperature does not influence the infection rate of schistosomiasis, but that total annual rainfall may.

Because of the way in which the MOD files and programs are set up, the user may query: "What is the infection rate (%) of schistosomiasis due to Schistosoma mansoni in man, in eastern Brazil, where (simultaneously) the

—— text continued page 8-18

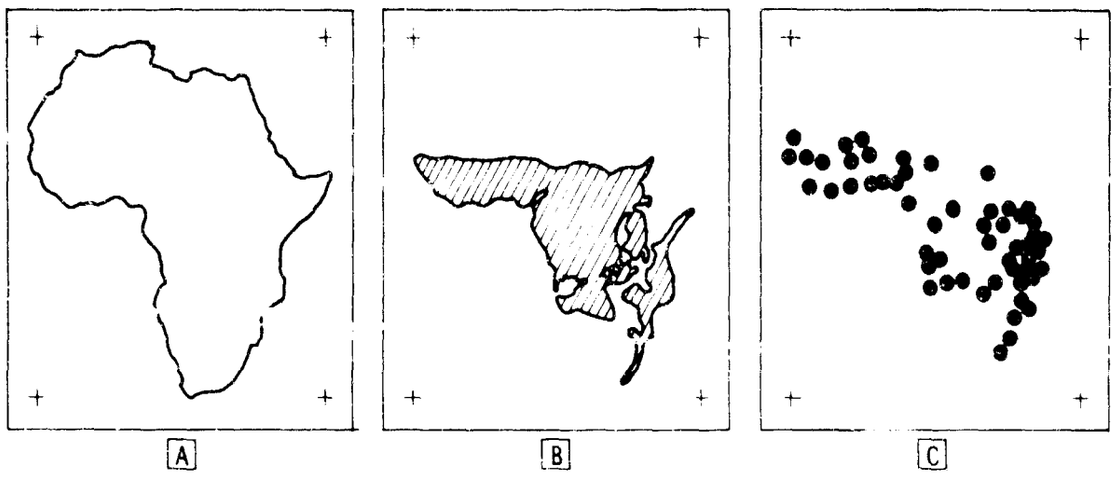
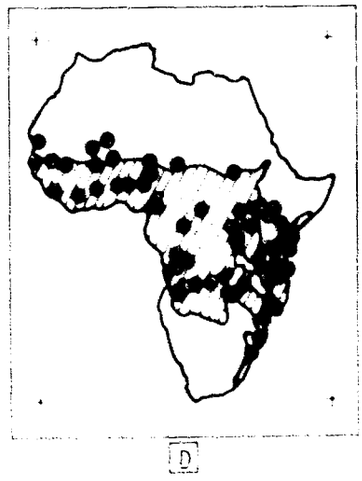


Figure 8-3 Data concerning Burkitt's tumor (Burkitt, 1962, p. 77-78; used by permission) recast into a MOD-like output form: A, a base map of Africa; B and C, maps which would be output by the MOD computer system to show: in B, areas (shaded) where these three conditions exist simultaneously -- altitude is under 5000 feet, seasonal mean temperature always exceeds 60°F, and total annual rainfall exceeds 20 inches, and, in C, occurrence (dots) of Burkitt's tumor. D, shows maps of A, B, and C overlaid to evaluate the extent of pattern match.



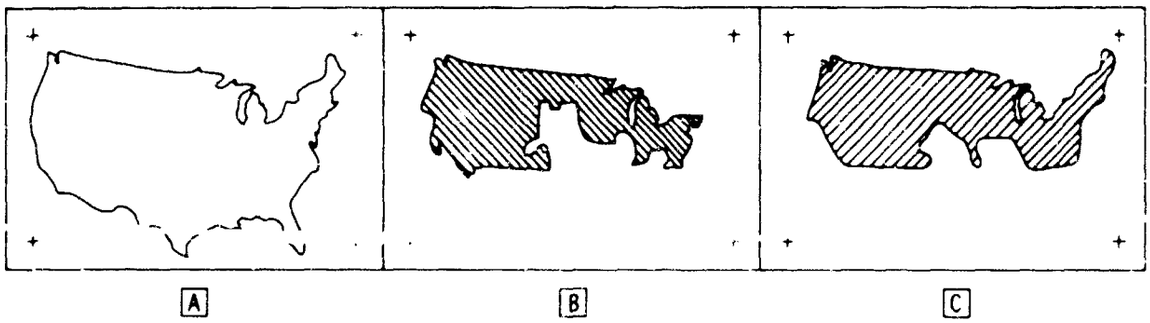
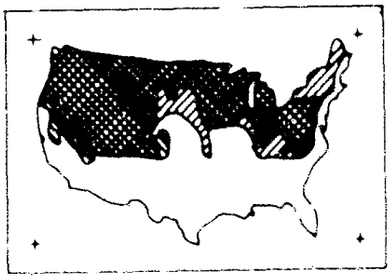


Figure 8-4 Data implying relationship between goiter and iodine content of drinking water -- Henschen, 1962, p. 190, about 1920 -- rearranged into MOD-like format. A, a base map of U.S.; B, a MOD map-like output showing areas (shaded) with iodine content of drinking water low (less than 0.23 parts per liter); C, a MOD map-like output showing areas (shaded) with goiter frequent (5 or more cases per 1000); D, maps of A, B, and C overlaid to show similarity of the distribution patterns of the two factors.



See fig. 3-4 p. 3-18.

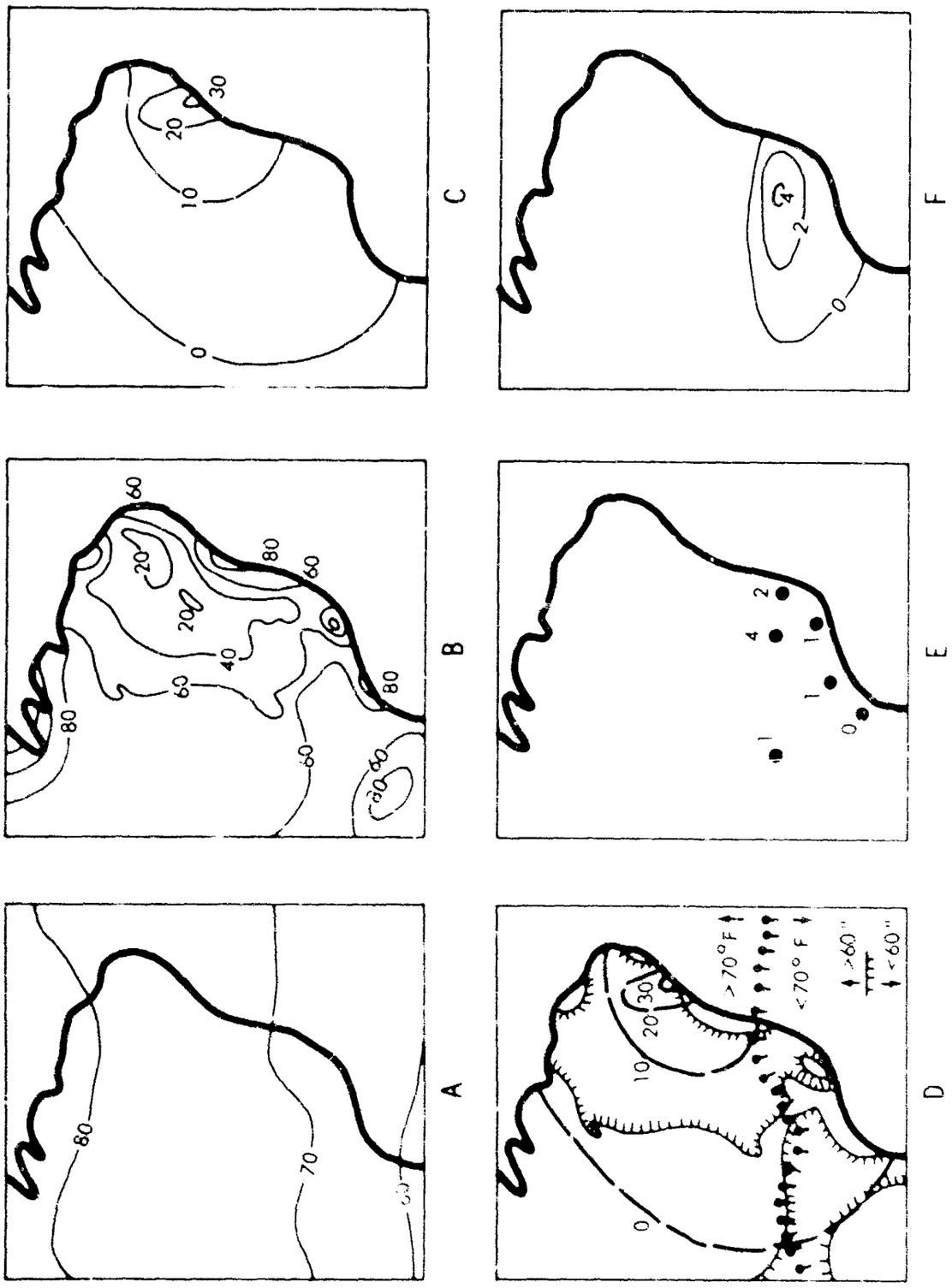


Figure 8-8

--- See opposing page

Figure 8-5 Maps showing distribution of temperature, rainfall, and schistosomiasis data in eastern Brazil: A, July normal temperature, ^{o}F, (Rand McNally, 1964, p. 11); B, Annual rainfall, inches, (Rand McNally, 1964, p. 97); C, Infection rate, %, of schistosomiasis mansoni in man, based upon data grouped by province (taken from Malek in May, 1961, p. 305-6), drawn manually by the MOD study team; D is a map made by overlaying A, B, and C, ($70^{o}F$ contour from A is shown as a dotted line; 60-inches (rainfall) contour from B is shown as a solid line; 0, 10, 20, and 30% (infection rate) contours from C are represented by dashed lines); E, Dot-type map showing data points that would have been retrieved and output by MOD system in response to query asking for the combined factors -- *infection rate of schistosomiasis mansoni in man where, simultaneously, July normal temperature is under 70°F and total annual rainfall exceeds 60 inches*; F, contour-type map drawn from data points of E.

annual rainfall exceeds 60 inches and the July normal temperature is less than 70°F?" In this case he would receive a single map (Fig. 8-5E,F), based only upon those data points which satisfied all the query conditions. This single map presents a distribution pattern which, when compared with the three separate maps (Fig. 8-5A,B,C), gives little insight into the over-all disease/environmental situation, nevertheless it describes a particular situation and does present potentially useful information.

The user could also request graphs (discussed earlier under Output Analysis) showing either schistosomiasis-rainfall-temperature or schistosomiasis-rainfall plots. But, again, it seems that relationships among the disease and environmental factors are most effectively shown (at least in the early stages of an investigation) by obtaining and comparing visually a group of separate maps, each displaying the geographic distribution of one simply-stated factor.

A fourth illustration makes use of some paleontological taxonomic and ecologic data (Ray, 1967) to explore a problem quite remote and far afield from the basic medical objectives of the MOD project. One of the reasons for this was to demonstrate that the MOD system is applicable to many areas other than the study of disease. Data which had actually been used in a study employing maps was recast into MOD-like output, then examined, leading to the same conclusions that were drawn by the original worker.

The first illustration (Fig. 8-6) shows a base map of the East coast and four maps, each showing only the geographic distribution of one environmental factor. (Each map was originally drawn on translucent overlay paper.)

Fossil walrus tusks of uncertain age, but possibly as old as several million years, have been found along the East coast from New England to Florida. If all these tusks are of Pleistocene (Ice Age) age and represent the living, cold-water species of walrus, it would seem that cold climate

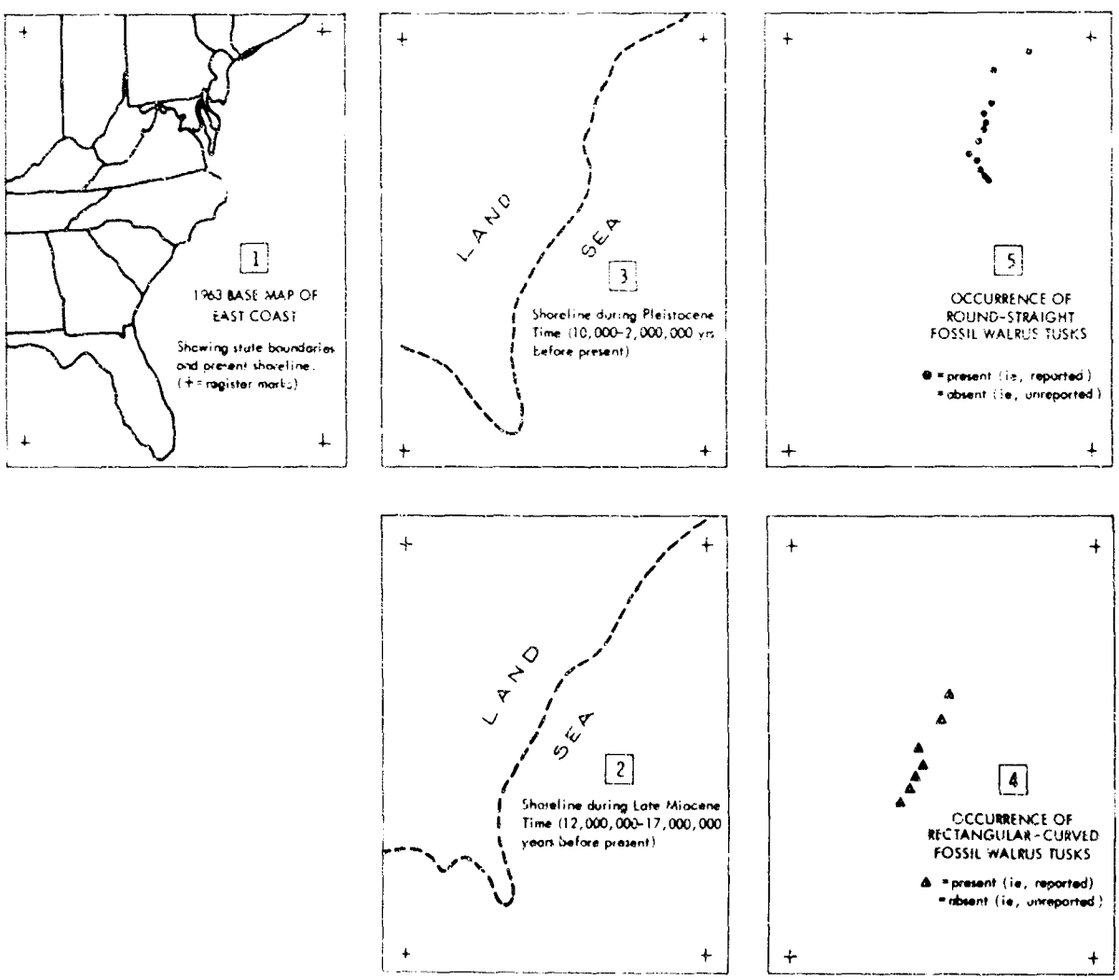


Figure 8-6 MOD-type maps, each showing the distribution of one environmental factor pertinent to study of fossil walruses; see Figure 8-7.

extended as far south as Florida. However, much other evidence indicates that Florida was only slightly cooler during the Ice Age than at present. Thus, there is a dilemma.

Close examination of the fossil walrus tusks shows that two morphologically different kinds occur, and that each occurs in a different geographic region (Fig. 8-6 and -4, -5).

Since the time when walruses became evolutionarily distinct from seals, the East Coast has been submerged by marine waters twice: once during the Late Miocene (12-17 million years ago), ard again during the Pleistocene (10,000-2,000,000 years ago). Studies of the deposits laid down during these submergences allow us to map the shorelines of these ancient seas (Fig. 8-6 and -2, -3).

When these shoreline maps are overlaid with the tusk-occurrence maps, it is immediately evident that the two kinds of fossil walrus tusks, in addition to being distinct morphologically and biogeographically, are also distinct paleoenvironmentally. (Remember that walruses are marine, not terrestrial animals.) The kind that is identical to the living cold-water walrus occurs predominantly in regions which were sea during Pleistocene time (but land during Late Miocene time); the other kind occurs in regions which were sea in Late Miocene time (but land during Pleistocene time), as shown in Fig. 8-7.

Thus, we resolve the apparent dilemma by conclusions, based upon our maps, that the more northern group of tusks are Pleistocene representatives of the living cold-water walrus species which ranged south only to North Carolina during the Ice Age, while the more southern tusks represent an earlier (late Miocene), now-extinct, warmer-water walrus species which ranged as far south as Florida.

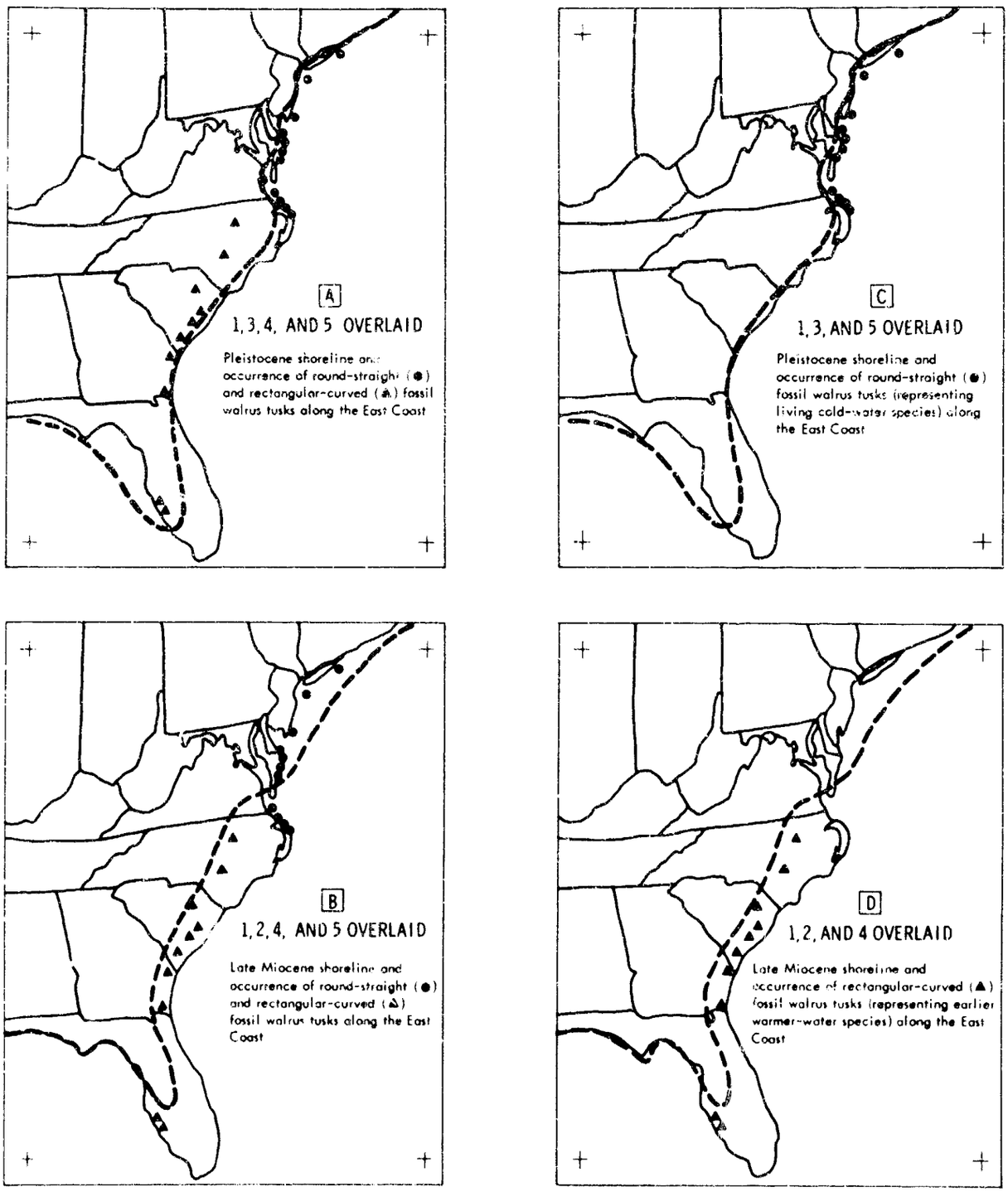


Figure 8-7 Overlaid combinations of appropriate MOD-type maps from Figure 8-6 to show how such maps could be used in resolving this paleontological problem (discussed in text). Maps A and B present all the data; Maps C and D present the data in a manner best suited to solving the problem.

9 General summary, conclusions, and recommendations

ABSTRACT - It would be difficult indeed to provide a meaningful short summary of the content of each of the preceeding eight sections, and that has not been attempted here. This section presents a general summary as a basis for drawing conclusions about the MOD effort and making specific recommendations.

Rene J. Dubos has pointed out (in his forward to "Attenuated Infection", 1960) that a very large amount of relevant information is available in the published literature which has remained virtually unnoticed because it has not been integrated in a meaningful pattern and correlated with the natural events of disease.

MAPPING OF DISEASE

9.1 GENERAL SUMMARY

The MOD project represents the first serious effort (to our knowledge) to develop a computerized system for mapping disease, coupled with a comprehensive data file of ecologic factors. Our goal was to provide a system whereby disease and environmental data could be manipulated together in an appropriate (geographic) location and time context, with direct computer (/line-printer or /plotter) output in the form of distribution maps, or block diagrams -- with supplemental narrative reports as required.

The implemented system would have a capacity for producing quickly, and easily, up-to-the moment maps that show the distribution patterns of diseases and causally related factors. In addition, the system would be an important research tool for those persons searching for new causal relationships, and/or attempting to predict changes in location patterns or incidence of disease. Obviously, an effective research method for linking contributing and precipitating factors with a given disease would aid in many ways our understanding of the etiology of disease. If the cause of the disease were unknown, it would be a means to define the communities with different incidences, and to analyze the differences between these communities as to environmental and other factors. In this way, factors of high correlation could be found, giving clues as to etiology and pointing to specific basic research which would be likely to define etiology and/or disclose methods of control.

In connection with the uses of the MOD system which we have envisioned, two excerpts from Professor A. Payne's "Statement on Epidemiology", made to the Executive Board of WHO (20 January 1966) are pertinent:

> *In the last analysis it is the ecology of an area which determines what diseases might become serious problems as conditions are changed in the process of development, or should any of a variety of agents be introduced. Knowledge of it therefore has a predictive value enabling one to foresee future dangers so that preventive action can be taken in good time.*

The second area of research which I would mention, involves the long-term development of ecological maps of the world, including the distribution of infectious agents, vectors, reservoirs and ecological conditions. I would emphasize that this is a long-term objective, but one which would lead to major advances in predictive epidemiology and communicable disease surveillance.

The immediate result of the MOD effort was envisioned as an *operational* computer system consisting of two major components:

- An information storage and retrieval system specifically designed for disease-environmental data
- A graphic output system that would manipulate retrieved data and present them in the form of maps (principally), block diagrams, graphs, and narrative reports.

An additional important result was to be the description of methods and techniques necessary to select, extract, evaluate, and preprocess "raw" narrative, tabular, and graphic data so that they could serve as effective input to the storage and retrieval system.

One of the last items that was to be produced for the MOD system was a user's manual. This was considered necessary because the MOD system will provide a unique capability, one with which the potential user will have had no experience.

Financial support was anticipated for a period of three years; it was provided for only two years and, as a result, the MOD system was not carried to the point of implementation. However, the system analysis and design have both been completed (with the exception of several aspects of system design that need further elaboration, but require that this be performed in the context of a partially implemented system). Furthermore, data characteristics have been extensively analyzed as to sources, limitations of the data, per se, and problems involved in preparing these data for

computer input. A method for structuring data has been designed and tested, and a comprehensive factor catalogue has been produced. In addition, through our analysis of maps and cartographic techniques, we have gained new insight into the characteristics of disease-environmental data that allow them to be mapped, and have developed data extraction forms reflecting these requirements.

* * *

Work on the MOD project has progressed to the point that feasibility is no longer a question. We have produced many disease-environmental maps as direct computer/plotter output, proving the validity of our hypotheses and demonstrating the adequacy of our data and our methods of data manipulation. We believe that we have developed the MOD system, not to completion, not to full satisfaction, not to implementation, but with every expectation of success.

9.2 CONCLUSIONS AND RECOMMENDATIONS

The conclusions and recommendations contained in this section reflect nearly three years of effort on the MOD project. To state these simply: (1) The (MOD) system for computerized mapping of disease-environmental data described herein is feasible; (2) This system would satisfy an important need for processing data to provide geographically oriented disease-environmental information; and (3) The MOD system should be implemented, and can be implemented -- given adequate time, effort, and financial support.

In particular, data-processing aspects should present no significant technical problems now that we have developed an effective method for structuring the data. Readily available computer science concepts, techniques, and equipment are adequate for this task, and no special difficulties are anticipated in producing the necessary programs. However, data-collecting aspects, especially extraction, will require a great effort (see Fig. 9-1, next page), and it is probably this phase, more than any other, that will limit use of the system.

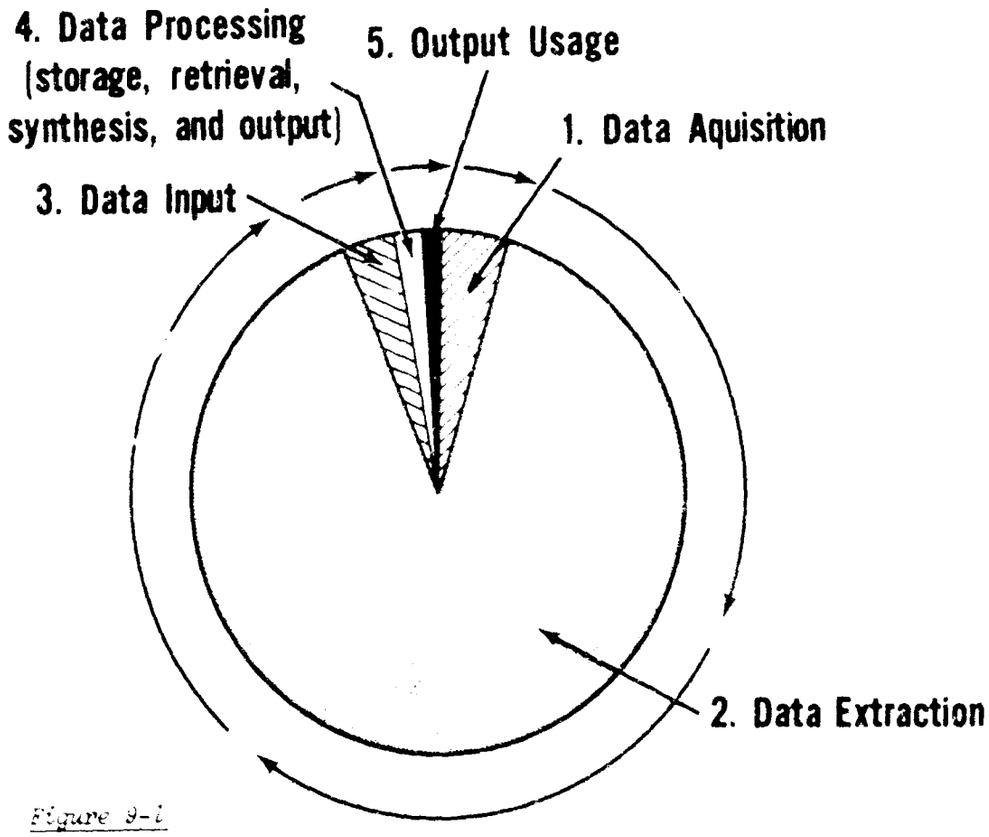


Figure 9-1

TOTAL MOD EFFORT

Many of the methods, techniques, and procedures that we have described can be implemented independently (to a limited extent), by imposing suitable restrictions on appropriate parts of the proposed MOD system. Ideally, the entire system should be implemented. If (funding) priorities do not allow this, we recommend that the system be implemented in part -- to whatever degree is permitted by available resources.

Implementation will require two or three competent computer programmers, working full-time for about one to one-and-a-half years under the direction of a computer-system analyst, in turn, supervised by a professional biomedical staff of two or three persons (who could also contribute to the

data collection/preprocessing effort). At least two or three biomedical professional, several semi-professional, and several clerical personnel will be required for the data-collecting-preprocessing efforts. Once the system becomes implemented, the number of data-collecting-preprocessing personnel required will depend entirely upon user requirements -- more specifically, the character and extent of the data base file required for effective response to the queries. A suggested table of staff organization is shown in Figure 9-2.

* * *

Requirements of the MOD system have been specified; feasibility has been proved; design has been accomplished. Implementation of the MOD system would provide a pwerful new tool to biomedical science. We conclude and strongly recommend that such a system be implemented as soon as possible.

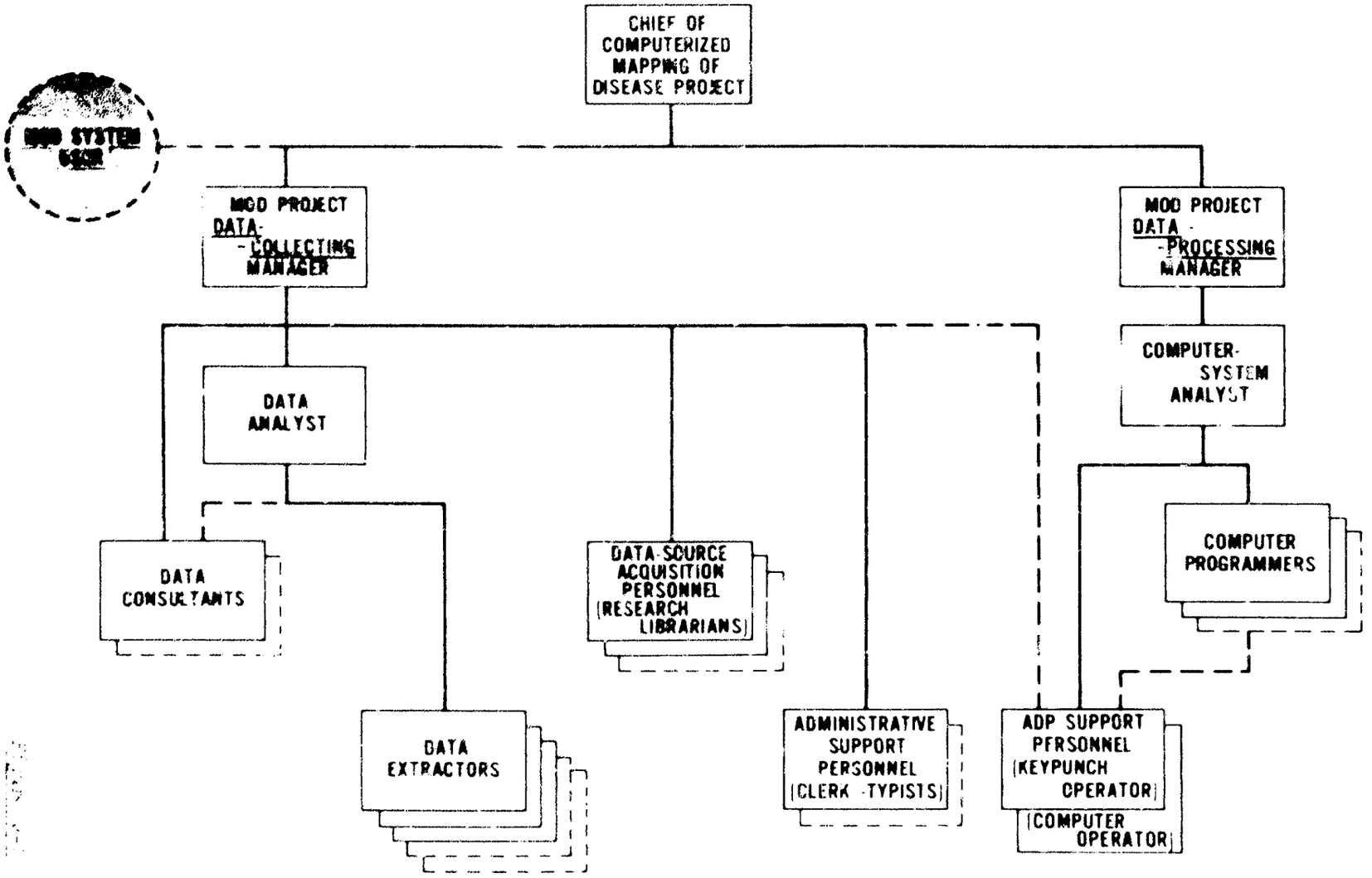


Figure 3-1. A suggested table of organization for personnel involved in operation of the fully implemented MOD system.

References cited

The list of references is divided into two parts: the first, References Cited -- the second, Selected Bibliography.

The items which appear among References Cited are not listed a second time.

References cited

American Geographical Society, World distribution of spirochetal diseases, 3., Leptospiroses: Atlas of Diseases, American Geographical Society, New York, pl. 17, 1955

Bick, K.F., and Johnson, G.H., Laboratory Manual in Earth Science, Crowell, New York 1967, 163 pp.

Bunge, W., Theoretical Geography; Gleerup, Lund, Sweden, 1962, 210 pp.

Burkitt, D., A tumor syndrome affecting children in tropical Africa, Postgrad. Med. J., v. 38 : 71-79, 1962

Datamation, Automatic Data Processing Glossary, Thompson, Greenwich, Conn., 1966, 62 pp.

De Paola, D., Pathology in Brazil -- past and present, International Pathology, v. 8 : 8-12, 1967

Digital Plotting Newsletter, November-December, Plotter perspective on U.S. population, California Computer Products Inc., Anaheim, Calif., p. 1, 1966

Dorn, H.F., A classification system for morbidity concepts, Public Health Reports, v. 72 : 1043-1048, 1957

Espenshade, E.G., Jr., editor, Goode's World Atlas, Rand McNally, Chicago, 1964, 288 pp.

Fisher, H.T., et al Introduction to synagraphic computer mapping -- Computer mapping of quantitative and qualitative information Introductory Correspondence Course, Laboratory for Computer Graphics, Harvard University, Cambridge, Mass., 1967

Harbaugh, J.W., A computer method for four-variable trend analysis illustrated by a study of oil-gravity variations in southeastern Kansas, State Geological Survey of Kansas, Bull. 171, 1964, 58 pp.

Henschen, F., The History and Geography of Diseases, Delacorte Press, New York, 1966, 344 pp. (English transl., Tate, J., 1966 -- orig. ed. 1962)

Howe, G.M., National Atlas of Disease Mortality in the United Kingdom, Royal Geographical Society, Thomas Nelson & Sons, London, 1963, 111 pp.

Kratchman, J., and Grahn, D , Relationships between the geologic environment and mortality from congenital malformation, TTD-8204, Biology & Medicine, Technical Information Service, U.S. Atomic Energy Commission, Washington, D.C., 1959, 20 pp.

References Cited

Learmonth, A.T.A., Health in the Indian subcontinent 1955-1964, Australian National University, Department of Geography, Occasional Paper 2, 1965, 80 pp.

_____, & Nichols, G.C., Maps of some standardised mortality ratios for Australia 1959-1963: Australian National University, Department of Geography, Occasional Paper 3, 1965, 35 pp.

Lobeck, A.K., Block diagrams and other graphic methods used in geology and geography, 2nd ed, Emerson-Trussell, Amherst, Mass., 1958

Malek, E.A., The ecology of schistosomiasis, in Studies in Disease Ecology, May, J.M., editor, Hafner, New York, 1961, pp. 261-327

May, J.M., editor, Atlas of Diseases, American Geographical Society, New York, 1950-55

Nature, Maps by machine, Nature, v. 213 : 1166-1167, 25 March, 1967

Nature, Instant maps, Nature, v. 214 : 230-231, 15 April, 1967

Ohman, H.L., Guidelines for constructing isopleth maps, Spec. Rept. S-2, Earth Sciences Division, U.S. Army Quartermaster Research & Engineering Center, Natick, Mass., (Proj. Ref. 1KO-25001-A129), 1963, 19 pp.

O'Leary, M., Lippert, R.H., and Spitz, O.T., FORTRAN IV and MAP program for computation and plotting of trend surfaces for degrees 1 through 6, State Geological Survey of Kansas, Computer Contribution 3, 1966, 48 pp.

Osborn, R.T., An automated procedure for producing contour charts, Informal Manuscript IM 67-4, U.S. Naval Oceanographic Office, Washington, D.C., 1967, 48 pp.

Petterssen, S., Introduction to Meteorology, 2nd ed., McGraw-Hill, New York, 1958, 327 pp.

Pfaltz, J.L., and Rosenfeld, A., Computer representation of planar regions by their skeletons, Communications of the Association for Computing Machinery, v. 10 : 119-125, 1967

Preston, F.W., and Harbaugh, J.W., BALGOL programs and geologic application for single and double Fourier series using IBM 7090/7094 computers, State Geological Survey of Kansas, Spec. Distrib. Pub. 24, 1965, 72 pp.

Ray, C.E., personal communication, 1967

Richardson, M.A. and Rollett, J.S., The Oxford Cartogra, Data Bank -- A feasibility study of accuracies, store sizes and operation time presented at the Third International Conference on Cartography, Amsterdam 17-22 April 1967

Robinson, A.H., Elements of Cartography, 2nd ed., John Wiley, New York, 1960, 343 pp.

Rodenwaldt, E., editor, World-Atlas of Epidemic Diseases, Falk-Verlag, Hamburg, Germany, 3 vols., 1952-58

Sippl, C.J., Computer Dictionary and Handbook, Howard W. Sams, Indianapolis, 1966, 766 pp.

Tobler, W.R., Notes on the analysis of geographical distributions, University of Michigan, Department of Geography, Michigan Inter-University Community of Mathematical Geographers Discussion Paper 8, part 2, 1966, 14 pp.

_____, personal communication, 1967

Selected bibliography

The Selected Bibliography is not meant to be exhaustive, but to serve as a guide.

The items which were listed among References Cited are not included here.

Selected bibliography

In addition to the "References Cited," we list here a group of selected references dealing with data processing, information storage and retrieval, epidemiology, medical ecology, and cartography that have helped to orient MOD project personnel. We believe that they may be of value to others concerned with the computerized mapping of disease and environmental data.

Abramson, N., Information Theory and Coding, McGraw-Hill, New York, 1963, 201 pp.

Abzug, I., Graphic data processing, Datamation, v. 11, no. 1 : 35-37, 1965

Adams, G.P., The use of a computer to calculate isodose information surrounding distributed gynaecological radium sources, Phys. Med. Biol., v. 9 : 533-540, 1964

Adams, S., and Taine, S., Searching the medical literature, J Amer. Med. Assoc., v. 188 : 251-254, 1964

Affel, H.A., System engineering, PR 7571, Auerbach Corp., (reprint from Internatl. Science & Tech.), 1964, 9 pp.

Ainsworth, G.C., Storage and retrieval of biological information, Nature, v. 191 : 12-14, 1961

American Association for the Advancement of Science, Symposium on information retrieval, Annual meeting of A.A.A.S., Washington, D.C., 1966

American Federation of Information Processing Societies, 1967 Spring Joint Computer Conference Proceedings, v. 30, 1967

American Geographical Society, Research catalogue of the American Geographical Society, Map supplement, Hall, Boston, 1962

American Rheumatism Association, Index of Rheumatology, v. 1, no. 23, 1965

Bibliography

Andrews, D., and Newman, S., Storage and retrieval of contents of technical literature -- Nonchemical information, Res. & Dev. Rept. 1, U.S. Patent Office, 13 pp, 1956

Andrews, R.D., and Ferris, D.H., Relationships between movement patterns of wild animals and the distribution of leptospirosis, Wildlife Management, v. 30 : 131-134, 1966

Anstey, R.L., A system for collation of environmental data, Res. Study Rept. RER-27, Regional Environments Research Branch, U.S. Army Quartermaster Research & Engineering Center, Natick, Mass., 1959, approx. 30 pp.

_____, Digitized environmental data processing, rev. ed., Res. Study Rept. RER-31, Regional Environments Research Branch, U.S.Army Quartermaster Research & Engineering Center, Natick, Mass., 1963, approx. 25 pp.

Arbib, M.A., Brains, Machines, and Mathematics, McGraw-Hill, New York, 1964, 152 pp.

Armstrong, R.W., Computer graphics in medical geography, Proc. Internatl. Geogr. Union, Latin Am. Regional Conf., v. 6 : 69-74, 1966

Artandi, S., Investigation of systems for the intellectual organization of information, Grant NSF GN99, Rutgers Univ., New Brunswick, N.J., 1964, 44 pp.

Asero, J.J., Find that fact, Army Information Digest, v. 21, no. 3 : 45-49, 1966

Atkins, H., editor, Proceedings of a One-Day Symposium on Progress in Medical Computing, Elliott Medical Automation Ltd., London, 1965

Auerbach Corp., Cancer Chemotherapy Abstracts, National Cancer Institute, Bethesda, Md., v. 6, no. 7-8, 1965

_____, A description of Auerbach Information Management System, Auerbach Corp., Philadelphia, 1966, approx. 50 pp.

Auerbach, I.L., The impact of information processing on mankind, PR 7603, Auerbach Corp., (reprint from Proc. of I.F.I.P. Congress 1962), 5 pp.

_____, Employment implications of automation, PR 7558, Auerbach Corp. (Senate Subcommittee on Employment & Manpower), 1963, 22 pp.

_____, The role of the systems designer/consultant, PR 7654, Auerbach Corp. (U.S. Army Automatic Data Processing Seminar for General Staff Officers), 1964, 10 pp.

_____, Tomorrow's outlook for EDP, PR 7623, Auerbach Corp., (reprint from J. of Data Management), 1964, 6 pp.

Aumen, W.C., A new map -- The numerical (digital) map (abstr.), ACSM-ASP 1966 Convention Prog., p. 1, 1966

Austin, C.J., Data processing aspects of MEDLARS, Bull. Med. Libr. Assn., v. 52 : 159-163, 1964

_____, The MEDLARS System -- An application report, Datamation, v. 10, no. 12 : 28-31, 1964

Baca, A., Prediction of the performance of a solution gas drive reservoir by Muskat's Equation, Kan. Geol. Surv., Comp. Contr. 8, 1967, 35 pp.

Bahn, R.C., et al, An information-retrieval system for research associated with the postmortem examination, Mayo Clin. Proc., v. 39 : 835-840, 1964

_____, Potential uses of a digital computer in the Section of Experimental & Anatomic Pathology, Mayo Clin. Proc., v 39 : 830-834, 1964

Baker, J.J., Scanning text with a 1401, Communications of the ACM, 1964

Baker, M.L., Automatic map compilation equipment for altitude measurements and orthophoto productions (abstr.), ACSM-ASP 1966 Convention Prog., p. 6, 1966

Barrer, L.A., Machine inventory of human pathological specimens, Proc. 4th IBM Med. Symp., pp. 105-108, 1962

Bartcher, R.L., FORTRAN IV program for estimation of cladistic relationships using the IBM 7040, Kan. Geol. Surv., Comp. Contr. 6, 1966, 54 pp.

Bartholomay, A F., The mathematical approach to the study of discrete biological events, Proc. 6th IBM Med. Symp., pp. 325-349, 1964

Baruch, J.J and Barnett, G.O., Joint venture at Massachusetts General, Datamation, v. 11, no. 12 : 29-33, 1965

Basile, A.S., The key to automated cartography -- A precision digital plotter system, Amer. Cong. on Surveying & Mapping, 1964 Regional Convention, Kansas City, Mo., 1964

Bassett, F.J., Machine information retrieval -- An annotated introduction and projection, Bull. Med. Libr. Assoc., v. 51 : p. 221-225, 1963

Battelle Memorial Institute, Directory of selected specialized information services Ad-Hoc Forum of Scientific and Technical Information Analysis Center Managers, Directors, and Professional Analysts, held at Battelle Memorial Institute, Columbus, Ohio, 124 pp., 1965

Baum, C., and Gorsuch, L., editors, Proceedings of the Second Symposium on Computer-Centered Data Base Systems, Tech. Mem. TM-2624/100/00, System Development Corp., Santa Monica, Calif., 1965, approx. 200 pp.

Bibliography

Becker, J., and Hayers, R.M., Information storage and retrieval -- Tools, elements, theories, John Wiley, N.Y., 1963, 448 pp.

Beckwith, H.M., Automatic map compilation system, Contract DA 44-009-eng-4596, FR, phase 1; RW-C129-202, Thompson Ramo Wooldridge Inc., Canoga Park, Calif., 1962, 83 pp.

Behrens, C., Computers and security, Science News, v. 91 : 532-533, 1967

_____, Computers that hear, Science News, v. 91 : 214, 1967

Bell, D.A., Information Theory and Its Engineering Applications, 3rd ed., Pitman, N.Y., 1962, 196 pp.

Bell Telephone Laboratories, Electronic graphics by computer, Science, v. 156 : 8, 1967

Benson-Lehner Corp., Computer Graphics (various issues), 1965-67

_____, LTE & STE FORTRAN IV PLOT subroutines, Publ. 548, Benson-Lehner Corp., Van Nuys, Calif., 1966, 55 pp.

_____, LTE & STE plotting systems, Publ. 551 DAB, Benson-Lehner Corp., Van Nuys, Calif., 1966, 48 pp.

Berman, M., Incomplete data and models, Proc. 6th IBM Med. Symp., pp. 647-653, 1964

Bernard, J., and Schilling, C.W., Accuracy of titles in describing content of biological sciences articles, Biological Sciences Communication Project, Amer. Inst. of Biol. Sci., Washington, D.C., 1963, 90 pp.

Bernstein, R.A., How Computers Work -- Operation Update, Workbook No. 1, Factory, Reader Service Dept., N.Y., 1962, 10 pp.

Bertram, S., and Beckwith, H.M., Type II Interim technical report for the automatic map compilation system, Rept. C129-109, Contract DA 44-009-eng-4596, Thompson Ramo Wooldridge Inc., Canoga Park, Calif., 1961, 43 pp.

Berul, L., Information storage and retrieval -- A state-of-the-art report, PR 7500-145, Auerbach Corp., Philadelphia, 1964, approx. 100 pp.

_____, Methodology and results of the DOD User Needs Survey, PR 7500-130, Auerbach Corp., Philadelphia, 1965, 24 pp.

Blakesley, R.G., The planning databank challenges the surveyor (abstr.), ACSM-ASP 1966 Convention Prog., p. 6, 1966

Blumenstock, D.I., The reliability factor in the drawing of isarithms, Annals of Assoc. of Amer. Geographers, v. 43, 1953

Bobrow, D.G., et al, The BBN-LISP System, Sci. Rept. 1, Contract AF 19(628)-5065, Bolt Beranek and Newman Inc., Cambridge, Mass., 1966, 82 pp.

Boggess, W.R., and Russell, R.L., Stream flow patterns on the Lake Glendale watershed in southern Illinois, Univ. of Illinois, Dept. of Forestry, Forestry Note 110, 1964, 5 pp.

Bolton, R.M., The potential of scanners in cartography (abstr.), ACSM-ASP 1966 Convention Prog., p. 7, 1966

Bonato, R.R., A general cross-classification program for digital computers, Behav. Sci., v. 6 : 347-357, 1961

Bonner, R.E., et al, DAP -- A diagnostic assistance program, Proc. 6th IBM Med. Symp., pp. 81-108, 1964

Borko, H., and Bernick, M.D., Toward the establishment of a computer-based classification system for scientific documentation, Tech. Memo. TM-1763, System Development Corp., Santa Monica, Calif., 1964, 49 pp.

Bosch, R., Account numbering and identification systems, PR 7686, Auerbach Corp., (Am. Bankers Assoc. Savings Bank Workshop), 1964, 12 pp.

Bousky, S., Scanning techniques for light modulation recording, Tech. Rept. AFAL-TR-66-188, Contract AF 33(615)-2632, Ampex Corp., Redwood City, Calif., 1966, 141 pp.

Brain, A.E., et al, Graphical data processing research study and experimental investigation, Quart. Progr. Repts. (1,2,3,4,5,6,7,8,15), Contract DA-36-039-sc-78343, Stanford Research Institute, Menlo Park, Calif., 1960-64

Breeding, K.J., et al, Order code for the film scanners of ILLIAC III, Rept. 176, Dept. of Computer Science, Univ. of Illinois, Urbana, 1965, 20 pp.

Breimann, R.J., Harvard picks Alexandria for computer map making, Evening Star, Washington, D.C., 1 Mar. 1966 issue

Briggs, L.I., and Pollack, H.N., Digital model of evaporite sedimentation, Science, v. 155 : 453-456, 1967

Brown, J., and Wagner, D., Subsystem for the digital coding and remote display of curved lines, Rept. IST-2900-219-F, Contract DA 36-039-sc-78801, Inst. Sci. & Tech., Univ. of Michigan, 1960, 51 pp.

Brunelle, R.H., Systems programs to accomodate biomedical research, pp. 127-140 of Stacy & Waxman, Computers in Biomedical Research, Academic Press, New York, 1965

Buck, C.P., et al, Investigation and study of graphic-semantic composing techniques, Rept. Eng. 695-615F, Contract AF 30(602)2091, Research Inst., Syracuse Univ., Syracuse, N.Y., 1961

Burck, G., The boundless age of the computer, in six parts, Fortune: March, 101-; April, 141-; May, 153-; June, 113-; August, 125-; October, 64-, 1964

Burkitt, D., A great pathological frontier, Postgrad. Med. J., v. 42 : 543-547, 1966

_____, and Hutt, M.S.R., An approach to geographic pathology in developing countries, International Pathology, v. 7, no. 1 : 106, 1966

_____, and Wright, D., Geographical and tribal distribution of the African lymphoma in Uganda, British Med. J., v. 1 : 569-573, 1966

Burzynski, E.F., UNAMACE -- Universal Automatic Map Compilation Equipment (abstr.), ACSM-ASP 1966 Convention Prog., p. 7, 1966

Cahn, J.N., Closing gaps in biological communications -- Need for a national voluntary plan for science information in the 70's, Fed. Proc., v. 22 : 993-1001, 1963

Cain, S.A., A definition of human ecology, Paper presented at Symp. on Human Ecology, 1966 Annual Meeting of American Association for Advancement of Science, Washington D.C., 1966

California Computer Products, Remot plotting, Bull. 123B, Calif. Comp. Prod. Inc., Anaheim, Calif., 1964

_____ _____, Digital Plotting Newsletter, various issues, 1964-1967

_____, Digital plotting systems, Bull. 175C, Calif. Comp. Prod. Inc., Anaheim, Calif., 1965, 16 pp.

Candarjs, G., et al, Presentation d'un code pour fiches perforees à tri électronique, Oncologia (Basel), v. 16 : 210-220, 1963

Cannon, H.L., Geochemical relations of zinc-bearing peat to the Lockport Dolomite, Orleans County, New York, U.S. Geol. Surv., Bull. 1000-D, pp. 119-185, 1955

_____, The development of botanical methods of prospecting for uranium on the Colorado Plateau, U.S. Geol. Surv., Bull. 1085-A, pp. 1-50, 1960

_____, The biogeochemistry of vanadium, Soil Science, v. 96 : 196-204, 1963

_____, and Bowles, J.M., Contamination of vegetation by tetraethyl lead, Science, v. 137 : 765-766, 1962

Carlson, W.M., A management information system designed by managers, Datamation, v. 13, no. 5 : 37-43, 1967

Carpenter, H.M., System for storage and retrieval of data from autopsies, Amer. J. Clin. Path., v. 38 : 449-467, 1962

_____, Data processing systems in pathology, Biomed. Sci. Instrum., v. 1 : 25-31, 1963

Casey, R.S., et al, Punched Cards -- Their Applications to Science and Industry, 2nd ed., Reinhold, New York, 1958, 697 pp.

Caster, W.O., Use of digital computer in study of eating habit patterns, Amer. J. Clin. Nutr., v. 10 : 98-106, 1962

Census Bureau, Map area computer, U.S Bureau of Census, 1964, 4 pp.

Chamberlin, W., The Round Earth on Flat Paper, National Geographic Society, Washington, D.C., 1947

Chayes, F., and Suzuki, Y., Geological contours and trend surface, J. Petrology, v. 4 : 307-312, 1963

Cheshier, R.G., Machine information search system, Bull. Med. Libr. Assn., v. 50 : 481-486 1962

Christopherson, W.M., and Mendez, W.M., A local geographic study of cervical cancer, International Pathology, v. 7, no. 4 : 103-105, 1966

Chung, C.S., Genetic analysis of human family and population data with use of digital computers, Proc. 3rd IBM Med. Symp., pp. 51-78, 1961

Clearinghouse for Federal Scientific & Technical Information (formerly Office of Technical Services), Selective Bibliographies, C.F.S.T.I. (O.T.S.), Washington, D.C., 1959-66

Coles, M.W., Applications of the electronic digital computer in nautical cartography (abstr.), ACSM-ASP 1966 Convention Prog., p. 13, 1966

Colilla, R.A., and Sams, B.H., Information structures for processing and retrieving: Communications of the ACM, 1962(?)

College of American Pathologists, Systematized Nomenclature of Pathology (SNOP), College of American Pathologists, Chicago, 1965, 439 pp.

Collins, G., Display software technology, Signal, July 1966 issue

Colner, B.J., Line-simulated map (abstr.), ACSM-ASP 1966 Convention Prog., p. 14, 1966

Commission Cooperative Technical Africa, Symposium on the survey needs of developing countries, Report of the Acting Secretary-General to the 18th session of the Commission, Section 9, 1963

Compendum Publications, Cardiovascular Compendium, v. 1, no. 5, 1966

Connelly, R.R., et al, End results in cancer of the lung - Comparison of male and female patients, J. Natl. Cancer Inst., v. 36 : 277-287, 1966

Connor, D.H., and Lunn, H.F., Buruli ulceration: Arch. Path. v. 81 : 183-199, 1966

Control Data Corporation, Control Data 3600 Computer System Reference Manual, Publ. 213b, Control Data Corp., Minneapolis, 1963, approx. 100 pp.

_____, Abstracts of available civil engineering applications, Control Data Corp., Minneapolis, 1965, 6 pp.

Cooley, J.C., A Primer of Formal Logic, Macmillan, New York, 1942, 378 pp.

Coons, S.A., Computer graphics and innovative engineering design, Datamation, v. 12, no. 5 : 32-34, 1966

_____, The uses of computers in technology, Scientific American, v. 215, no. 3 : 176-188, 1966

Coppock, J.T., Electronic data processing in geographical research, The Professional Geographer, v. 14, no. 4 : 1-4, 1962

Corbin, H.S., A survey of CRT display consoles, Control Engineering, Reuben H. Donnelly Corp., New York, 1965, 8 pp.

Creighton, R., The Pacific Project Data System - A tool for the utilization of bird data, Information Systems Division, Smithsonian Inst., Washington, D.C., 1966

_____, and Humphrey, P.S., Application of Automatic Data Processing to the study of seabirds, Dept. of Vertebrate Zoology, Natural History Museum, Smithsonian Inst., Washington, D.C., 1966

Cude, W.C., Automation in mapping, Surveying & Mapping, v. 22 : 413-436, 1962

Cunningham, B.T., Coding of pathologic diagnoses at the Armed Forces Institute of Pathology, Amer. J. Clin. Path., v. 25 : 1181-1182, 1955

Cutler, S.J., Trends in cancer therapy and patient survival, 1940 to 1959, Natl. Inst. Health, Natl. Cancer Inst., Bethesda, Md., pp. 745-759, 1965

_____, and Latourette, H.B., A national cooperative program for the evaluation of end results in cancer, J. Nat. Cancer Inst., v. 22 : 633-646, 1959

Datamation, Data transmission systems, Datamation, v. 11, no. 12 : 51-53, 1965

Davis, J.C., Application of response-surface analysis to sedimentary petrology, Kan. Geol. Surv., Computer Contrib. 12, pp. 57-62, 1967

_____, and Sampson, R.J., FORTRAN II program for multivariate discriminant analysis using an IBM 1620 computer, Kan. Geol. Surv., Comp. Contr. 4, 1966, 8 pp.

Dayhoff, M.O., A contour-map program for X-ray crystallography, Communications of the ACM, v. 6 : 620-622, 1963

DeMeter, E.R., The influence of automation on mapping requirements and techniques (abstr.): ACSM-ASP 1966 Convention Prog., p. 11, 1966

Dempsey, J.R., A generalized two-dimensional regression procedure, Kan. Geol. Surv., Comp. Contrib. 2, 1966, 12 pp.

Department of Defense, Disease and Injury Codes, U.S. Army Tech. Bull., TB MED 15, 1963, approx. 650 pp.

Department of the Army, Cartographic Aerial Photography, Tech. Man. TM 5-243, U.S. Govt. Printing Office, Washington, D.C., 1964, 63 pp.

Derrick, E.H., et al, Epidemiological observations on leptospirosis in north Queensland, Australasian Annals of Medicine, v. 3, no. 2 : 85-97, 1954

Desautels, A.V., Automatic point marking and measuring instrument test results (abstr.), ACSM-ASP 1966 Convention Prog, p. 12, 1966

Dietrich, E.V., Machine retrieval of pharmacological data, Science, v. 132 : p. 1556-1557, 1960

Digital Equipment Corp., Computers in oceanography, Publ. G-8260, Digital Eqpt. Corp., Maynard, Mass., 1965, 8 pp.

_____, The Digital Logic Handbook, 1966-67 edition; Digital Eqpt. Corp., Maynard, Mass., 1966, 330 pp.

Dillon, E.L., and Nichols, C.W., Handling of statistical well data by computer, Amer. Assoc. Petrol. Geol., v. 49 : 1520-1531, 1965

Dingman, H.F., Computer analysis of psychological and psychiatric data, pp. 331-350 of Stacy & Waxman, Computers in Biomedical Research, 1965

Dixon, P., Decision tables and their application, PR 7568, Auerbach Corp., (reprint from Computers and Automation, v. 13, no. 4), 1964, 8 pp.

Dixon, P.J., and Sable, J., DM-1 - A generalized data management system, Spring Joint Computer Conference, AFIPS Conf. Proc., v. 30 : 185-198, 1967

Documentation Inc., Actual and potential association of ideas in information systems, Tech. Rept. 3, Contract Nonr-1305(00) Documentation Inc., Washington, D.C., 1954, 6 pp.

Dodd, J.R., Cain, J.A., and Bugh, J.E., Apparently significant contour patterns demonstrated with random data, J. Geol. Ed., v. 13 : 109-112, 1965

Dorn, H.F., and Cutler, S.J., Morbidity from cancer in the United States, Publ. Hlth. Mon. 56, Natl. Cancer Inst., Bethesda, Md., 1959

Dotson, J.C., editor, Short course on computers and computer applications in the mineral industry, College of Mines, Univ. of Arizona, 1961

_____, and Peters, W., editors, Computers and computer applications in mining and exploration, Univ. of Arizona, Tucson, Ariz., 1961

Dreyfus, R.H., et al, Tulane Information Processing System, Version II including MEDITRAN, Monogr. 3, Computer Science Series, Tulane Univ., 1966, 45 pp.

Drosness, D.L., et al, The application of computer graphics to patient origin study techniques, Publ. Health Rept., v. 80 : 33-40, 1965

Duncan, O.D., Cuzzort, R.P., and Duncan, B., Statistical Geography -- Problems in Analyzing Areal Data, Free Press of Glencoe, Glencoe, Ill., 1961, 191 pp.

Dunlap and Associates, The Department of the Army ENVANAL System, 11th Progr. Rept., Contract DA44-109-qm-1561, Dunlap & Associates Inc., Stamford, Conn., 1955, 25 pp.

_____, ENVANAL -- Field test of system on Operation MOOSEHORN, Progr. Rept. 2-2, Contract No. DA19-129-qm-390, Dunlap & Associates Inc., Stamford, Conn., 1956, 163 pp.

_____, Project ENVANAL, Final Report on Research Phase, Contract DA19-129-qm-390, Dunlap & Associates Inc., Stamford, Conn., 1956, 110 pp.

Eberhart, J., About the systems system, Science News, v. 91 : 19, 1967

Eden, M., Pattern analysis, Proc. 3rd IBM Med. Symp., pp. 215-232, 1961

Edmundson, H.P., Automatic abstracting; Final Rept. C107-3U1 (RADC-TDR-63-93), Contract AF 30(602)2223, TRW Computers Co. Canoga Park, Calif., 1963, 91 pp.

Einhorn, S.J., Reliability prediction for repairable redundant systems, Proc. of the IEEE, v. 51 : 312-317, 1963

Empey, S.L., Computer applications in medicine and the biological sciences bibliography, Rept. SP-1025, System Development Corp., Santa Monica, Calif., 1962, 38 pp.

Eubanks, F.R., and Baker, G.T., Array Research, Automated Mapping Systems, Spec. Rept. 11, Contract AF 33(657)-12747, Texas Instruments Inc., Dallas, Tex., 1966, 40 pp.

Evans, D.C., Computer logic and memory: Scientific American, v. 215, no. 3 : 74-85, 1966

Fairchild Camera & Instrument Corp., Viewer, still picture, Final Development Rept. (SME-AG-3; RADC TR 58-160), Contract AF 30(602)1727, Fairchild Cam. & Inst. Corp., Syosset, N.Y., 1958, 18 pp.

Fano, R.M., and Corbató, F.J., Time-sharing on computers, Scientific American, v. 215, no. 3 : 128-140, 1966

Favret, A.G., Introduction to Digital Computer Applications, Reinhold, New York, 1965, 246 pp.

Feidelman, L., A survey of the character recognition field, PR 7593, Auerbach Corp., 1966, 25 pp.

Fleischer, M., Fluoride content of ground water in the conterminous United States, U.S. Geol. Surv., Misc. Geol. Investig., Map I-387, 1962

FMA Inc., The File-Search System, general information manual, FMA Inc., Washington, D.C., 1964, 28 pp.

Fox, W.T., FORTRAN IV program for vector trend analyses of directional data, Kan. Geol. Surv., Comp. Contr. 11, 1967, 36 pp.

Fuchs, A., Geography of eye disease, Notring der Wissenschaftlichen Verbände Österreichs, Wien, 1962, 162 pp.

Garfield, E., Citation indexes for science -- A new dimension in documentation through association of ideas, Science, v. 122 : 108-111, 1955

Garfinkel, D., Digital computer simulation of ecological systems, Nature, v. 194 : 856-857, 1962

_____ _____, Programmed methods for printer graphical output, Communications of the A.C.M., v. 5 : 477-479, 1962

_____, Simulation of ecological systems, pp. 205-216 of Stacy and Waxman, Computers in Biomedical Research, Academic Press, New York, 1965

_____, et al, Computer simulation and analysis of simple ecological systems, Ann. N.Y. Acad. Sci., v. 115 : 943-951, 1964

Garrett, P., Classification system for any data banking (information storage and retrieval) process, Res. Rept. 59-6, Contract Nonr-2666(00), Benson-Lehner Corp., Santa Monica, Calif., 1959, 11 pp.

Garvey, W.D., and Griffith, B.C., Scientific communication as a social system, Science, v. 157 : 1011-1016, 1967

Gaul, R.D., Instrumentation and data handling system for environmental studies off Panama City, Fla., Ref. no. 62-IT, Contract Nonr-211904, Texas A. & M. College, College Station, Tex., 1962, 6 pp.

Gerard, R.W., Quantitation in biology, Proc. 4th IBM Med. Symp., pp. 29-48, 1962

Giammo, T.P., A mathematical method for the automatic scaling of a function, J. Assn. for Computing Machinery, v. 11, no. 1 : 79-83, 1964

Gilbert, E.N., Information theory after 18 years, Science, v. 152 : 320-325, 1966

Gilles, H.M., Akufo -- An Environmental Study of a Nigerian Village Community, Ibadan Univ. Press, Ibadan, Nigeria, 1964, 80 pp.

Gittelsohn, A.M., et al, Tabulation of vital records by computer, Public Health Rept., v. 79 : 895-904, 1964

Gordon, B.L., editor, Current Medical Terminology (CMT), 3rd ed., Amer. Med. Assoc., Chicago, 1966, 969 pp.

Gosden, J.A., Estimating computer performance, Computer Journal, v. 5, no. 4 : 276-283, 1964

_____, and Sisson, R.L., Standardized comparisons of computer performance, PR 0624, Auerbach Corp., (reprint from Proc. of IFIP Congress 1962, pp. 57-61), 1962

Gray, H., et al, Information retrieval and the design of more intelligent machines, Final Rept. for 1 May 58 - 30 Jun 59, Project ADAR, Task E, Contract DA 36-039-sc-75047, Moore School of Elect. Eng., Univ. Penn., 1959, 216 pp.

_____, and Parker, E., Information retrieval and the design of more intelligent machines, Final Rept. for 1 Jul 59 - 30 June 60, Task E, Contract DA 36-039-sc-75047, Moore School of Elect. Eng., Univ. Penn., 1960, 77 pp.

Greanias, E.C., The computer in medicine, Datamation, v. 11, no. 12 : 25-28, 1965

Green, M., Design factors for data transmission systems, PR 7651, Auerbach Corp., (1964 Internatl. Symposium on Global Communications), 1964, 19 pp.

Greenberger, M., The uses of computers in organizations, Scientific American, v. 215, no. 3 : 192-202, 1966

Greenly, J.F., Standardization of typewriter fonts for automatic reading, Rept. for U.S. Air Force (RADC-TR-65-523), General Precision's Link Group, Binghamton, N.Y., 1966, 53 pp.

Griffith, W.H., A study of the rationale and techniques for long-range technological forecasting in the biological and medical sciences, Rept. for Life Sciences Div. of Army Research Office, Contract DA-49-092-ARO-9, Fed. Amer. Soc. for Exper. Biol., Washington, D.C., 1964, 52 pp.

Griffiths, J.C., Statistical approach to the study of potential oil reservoir sandstones, pp. 637-668 of Parks, G.A., editor Computers in the mineral industries, Stanford Univ. Press, Palo Alto, 1964

Hambleton, W.W., New dimensions for mineral resources studies, Kan. Geol. Surv., Spec. Dist. Pub. 31, 1966, 7 pp.

Hammer, C., Software considerations for management information systems, Montreal Chapter, Data Processing Mgmt. Assoc., Montreal, Quebec, Canada; (reprint of invited paper given 1 Jul 65 at DPMA Internatl. Data Processing Conf., Philadelphia), 1965, 29 pp.

Harbaugh, J.W., Direct printing of computer maps of facies data by computer (abstr.), Amer. Assoc. Petrol. Geol. Bull., v. 46 : 268, 1962

_____, Trend-surface mapping of hydrodynamic oil traps with the IBM 7090/7094 computer, Quart. Colo. Sch. Mines, v. 59 : 557-578, 1964

_____, Mathematical simulation of marine sedimentation with IBM 7090/7094 computers, Kan. Geol. Surv., Comp. Contr. 1, 1966, 52 pp.

_____, and Demirmen, F., Application of factor analysis to petrologic variations of Americus Limestone (Lower Permian), Kansas and Oklahoma, Kan. Geol. Surv., Spec. Dist. Publ. 15, 1964, 40 pp.

_____, and Preston, F.W., Fourier series analysis in geology, Short Course and Symp. on Computers and Comp. Applications in Mining and Exploration, Univ. of Arizona, v. 1 : R-1 - R-46, 1965

_____, and Wahlstedt, W.J., FORTRAN IV program for mathematical simulation of marine sedimentation with IBM 7040 or 7094 computers, Kan. Geol. Surv., Comp. Contr. 9, 1967, 40 pp.

Harris, J.N., and Madle, E.J., The ACM-52 automatic clutter mapper and preliminary experimental results, Tech. Rept. 206, Contract AF 19(604)5200, Lincoln Lab., Mass. Inst. of Tech., 1959, 52 pp.

Hastings, A.D., Atlas of Arctic Environment, RER-33, Regional Environments Research Branch, U.S. Army Quartermaster Research and Engineering Center, Natick, Mass., 1961, 22 pp.

Hathaway, J.P., How BUSHIPS automatically stores and retrieves documents, Paper presented at NARS Symposium on Office Information Retrieval, 1962, 6 pp.

Hayes, O.B., et al, Computers in epidemiologic dietary studies, J. Amer. Diet. Assn., v. 44 : 456-460, 1964

Head, R.V., Management information systems -- A critical appraisal, Datamation, v. 13, no. 5 : 22-27, 1967

Helava, U.V., A family of photogrammetric systems (abstr.), ACSM-ASP 1966 Convention Prog., p. 18, 1966

Hershey, A.V., The plotting of maps on a CRT printer, Rept. 1844, U.S. Naval Weapons Lab., Dahlgren, Va., 1963, 79 pp.

Hienz, H.A., et al, Zur Frage der Dokumentation in der pathologischen Anatomie, Medizinische Dokumentation, v. 5 : 10-12, 1961

Hobson, R.D., FORTRAN IV programs to determine surface roughness in topography for the CDC 3400 computer, Kan. Geol. Surv., Comp. Centr. 14, 1967, 28 pp.

Hodes, L., Machine processing of line drawings, Rept. (LL-54G-0028), Contract AF 19(604)7400, Lincoln Lab., Mass. Inst. of Tech., 1961, 15 pp.

Hoffman, J., Digitizing bathymetric data aboard ship for processing by automation (abstr.), ACSM-ASP 1966 Convention Prog., p. 23, 1966

Hopps, H.C., and Gabrieli, E.R., A new look at "normal" values in clinical pathology (editorial), International Pathology, v. 9, no. 1 : 10-11, 1968

Hopps, H. C., Data versus information (editorial), International Pathology, v. 8, no. 2 : 39-40, 1967

Hopps, H. C., Information -- A problem in geographic pathology (editorial), International Pathology, v. 8, no. 1 : 14-15, 1967

Horowitz, A.S., Discussion and description of cataloging system for the paleontological collections of the Indiana Univ. Dept. of Geology and the Ind. Geological Survey, mimeographed paper, 1964

Hough, P.V., General purpose visual input for a computer, Ann. N.Y. Acad. Sci., v. 99 : 323-334, 1962

Houston, N. and Wall, E., The distribution of term usage in manipulative indexes, Amer. Documentation, v. 15 : 105-114, 1964

Humphrey, P.S., An ecological survey of the Central Pacific, Smithsonian Year 1965 (Smiths. Inst., Ann. Rept. for year ended 30 Jun 65), pp. 24-30, 1965

Huntington, E. and Shaw, E.B., Principles of Human Geography, 6th ed., John Wiley, New York, 1951

International Business Machines Corp., Numerical code for states, counties, and cities of the United States, Int. Bus. Mach. Corp., New York, 1952, 81 pp.

_____, Computer Set AN/GSQ-16(XW-1), v. 1,2, and 6, Final Rept. (RADC-TR-59-110), Contract AF 30(602)1823, Int. Bus. Mach. Corp., Yorktown Heights, N. Y., 1959

_____, Proceedings of the 1st and 2nd IBM Medical Symposia, Int. Bus. Mach. Corp., Yorktown Heights, N. Y., 1961, 427 pp.

_____, Graphic composing techniques, Final Rept. (RADC TDR 61-310), Contract AF 30(602)2527, Thomas J. Watson Research Center, Int. Bus. Mach. Corp., Yorktown Heights, N.Y., 1962, 67 pp.

_____, Proceedings of the 3rd IBM Medical Symposium, Int. Bus. Mach. Corp., Yorktown Heights, N. Y., 1962, 575 pp.

_____, Proceedings of the 4th IBM Medical Symposium, Int. Bus. Mach. Corp., Yorktown Heights, N.Y., 1962, 512 pp.

_____, Proceedings of the 5th IBM Medical Symposium, Int. Bus. Mach. Corp., Yorktown Heights, N.Y., 1963, 502 pp.

_____, Numerical surface techniques and contour map plotting, Publ. E. 20-0117-0, Int. Bus. Mach. Corp., Yorktown Heights, N. Y., 1964

_____, Proceedings of the 6th IBM Medical Symposium, Int. Bus. Mach. Corp., Yorktown Heights, N.Y., 1964, 653 pp.

_____, Turning time ahead, Computing Report, v. 2, no. 3 : 8-12, 1966

Jackson, V.N., The multipoint system of digital structural analysis (abstr.), ACSM-ASP 1966 Convention Prog., p. 19, 1966

Jahn, T.L., The use of computers in systematics, J. Parasit., v. 48 : 656-663, 1962

James, W.R., FORTRAN IV program using double Fourier series for surface fitting of irregularly spaced data, Kan. Geol. Surv., Comp. Contr. 5, 1966, 19 pp.

Janaske, P.C., editor, Information handling and science information -- a selected bibliography 1957-1961, Biological Sciences Communication Project, Amer. Inst. of Biol. Sci., Washington, D.C., 1962, approx. 100 pp.

Jenks, G.F. and Brown, D.A., Three-dimensional map construction, Science, v. 154 : 857-864, 1966

Joint Council Subcommittee on Cerebrovascular Disease, Cerebrovascular Bibliography, v. 5, no. 2, 1965

Journal of the American Medical Association, The role of computers in modern medicine (editorial), J. Amer. Med. Assoc., v. 196 : 196-197, 1966

Juenemann, H.G., The design of a data-processing center for biological data, Ann. N.Y. Acad. Sci., v. 115 : 547-558, 1964

Kaesler, R.L., et al, FORTRAN II program for coefficient of association (MATCH-COEFF) using an IBM 1620 computer, Kan. Geol. Surv., Spec. Dist. Pub. 4, 1963, 9 pp.

Kao, R.C., The use of computers in the processing of geographic information, Geographical Review, v. 53 : 530-547, 1963

Kaufman, W.C., Standardization of symbols and units for environmental research, AMRL-TR-66-115, U.S. Air Force Aerospace Medical Research Laboratories, Wright Patterson A.F.B., Ohio, 1966, 4 pp.

Kay, M., et al, The Catalog Input/Output System, Memorandum RM-4540-PR, Rand Corp., Santa Monica, Calif., 1966, 71 pp.

Kellaway, G.P., Map projections, Methuen, London, 1949

Kent, A., and Perry, J.W., The storage and retrieval of nonnumerical data in large and complex documentation systems, Tech. Note 6, Contract AF 49(638)357, Center for Documentation and Communication Research, Western Reserve Univ., 27 pp.

Keyser, S.J., Advanced language processing procedures, ESD-TDR-63-620, Directorate of Computers, U.S. Air Force Systems Command, Bedford, Mass., 1963, 20 pp.

Kiefer, J., et al, Channels with arbitrarily varying channel probability functions, Information and Control, v. 5 : 44-54, 1962

King, G., et al, Automation and the Library of Congress, U.S. Govt. Printing Office, Washington, D.C., 1964, 88 pp.

Kleinmuntz, B., Clinical information processing -- Problem-solving strategies, Datamation, v. 11, no. 12 : 41-49, 1965

Klingbiel, P.H., Language oriented retrieval systems, Armed Services Technical Information Agency, Arlington, Va., 1962, 100 pp.

Korein, J., et al, Computer processing of medical data by variable field length format, J. Amer. Med. Assoc., v. 186 : 132-138, 1963

_____, Computer processing of medical data by variable-field-length format, J. Amer. Med. Assoc., v. 196 : 132-145, 1966

Krumbein, W.C., Trend surface analysis of contour-type maps with irregular control-point spacing, J. Geophys. Rsch., v. 64 : 823-834, 1959

_____, Computer analysis of stratigraphic maps (abstr.), Amer. Assoc. Petrol. Geol. Bull., v. 46 : 270, 1962

_____, The computer in geology: Science, v. 136 : 1087-1092, 1962

_____, FORTRAN IV computer programs for Markov chain experiments in geology, Kan. Geol. Surv., Comp. Contr. 13, 1967, 38 pp.

_____, and Imbrie, J., Stratigraphic factor maps, Amer. Assoc. Petrol. Geol. Bull., v. 47 : 698-701, 1963

_____, and Sloss, L.L., High-speed digital computers in stratigraphic and facies analysis, Amer. Assoc. Petrol. Geol. Bull., v. 42 : 2650-2669, 1958

_____, and _____, Stratigraphy and Sedimentation, 2nd ed., Freeman, San Francisco, 1963, 660 pp.

Lamson, B.G., and Dimsdale, B., A natural language information retrieval system, mimeographed paper, Pub. Hlth. Svc., Grant HM 00300-01, Univ. Cal., Los Angeles, 1966, 13 pp.

Latham, J.P., Possible applications of electronic scanning and computer devices to the analysis of geographic phenomena, Tech. Rept. 1, Contract Nonr-551(29), Wharton School of Finance and Commerce, Univ. Penn., 1959, 27 pp.

_____, A study of the application of electronic scanning and computer devices to the analysis of geographic phenomena, Final Rept., Contract Nonr-551(29), Wharton School of Finance and Commerce, Univ. Penn., 1959, 6 pp.

Ledley, R.S., and Lusted, L.B., Reasoning foundations of medical diagnosis, Science, v. 130 : 9-21, 1959

_____, and Ruddle, F.H., Chromosome analysis by computer, Scientific American, v. 214, no. 4 : 40-48, 1966

Leeds, H.D., and Weinberg, G.M., Computer Programming Fundamentals, McGraw-Hill, New York, 1961, 368 pp.

Leibholz, S.W., Introduction to system effectiveness evaluation, PR 7500-057, Auerbach Corp., (presented to George Washington Univ. School of Engineering & Applied Schience Center for Measurement Schience), 1965

Levy, W.A., Techniques for digital representation of terrain (abstr.), ACSM-ASP 1966 Convention Prog., 1965, pp. 25

Lewis, C.I., and Langford, C.H., Symbolic Logic, 2nd ed., Dover, New York, 1959, 518 pp.

Lewis, R.F., KWIC -- Is it quick?, Bull. Med. Libr. Assoc., v. 52 : 142-147, 1964

Liberman, E., Descriptors and computer codes used in Naval Ordnance Laboratory Library Retrieval Program, TR 64-20, U.S. Naval Ordnance Lab., White Oak, Md., 1964, 228 pp.

_____, and Stevens, H.L., Tables of four-letter computer codes used in library retrieval program, Rept. NOLTR 62-50, U.S. Naval Ordnance Lab., White Oak, Md., 1962, 680 pp.

Licht, S., editor, Medical Climatology, Elizabeth Licht, New Haven, Conn., 1964, 753 pp.

Light, D.L., Ranger mapping by analytical topographic compilation (abstr.), ACSM-ASP 1966 Convention Prog., pp. 25-26, 1966

Lipetz, B.A., Information storage and retrieval, Scientific American, v. 215, no. 3 : 224-242, 1966

Lipkin, M., and Woodbury, M.A., Coding of medical case history data for computer analysis, Communications of the ACM, 1962

Livingstone, F.C., Computer diagnosis, Science News, v. 91 : 558, 1967

Loomis, R.G., Boundary networks, Communications of the ACM, v. 8 : 44-48, 1965

Losee, F.L., Trace element variables related to oral health, pp. 41-54 of Environmental Variables in Oral Disease, Amer. Assoc. Advanc. Sci., 1966

Lunin, M., Coordinate indexing for information retrieval in an oral pathology department, Oral Surg., v. 18 : 484-493, 1964

Macgraith, B., Exotic Diseases in Practice, William Heinemann Medical Books, London, 1965, pp. 361

Bibliography

Manson, V., and Imbrie, J., FORTRAN program for factor and vector analysis of geologic data using an IBM 7090 or 7094/1401 computer system, Kan. Geol. Surv., Spec. Dist. Publ. 13, 1964

Marden, E.C., and Koller, H.R., Survey of computer programs for chemical information searching, Tech. Note 85, U.S. National Bureau of Standards, Washington, D. C., 1961, 87 pp.

Mason, E.E., and Bulgren, W.G., Computer Applications in Medicine, C.C. Thomas, Springfield, Ill., 1963

Mazfield, M., et al, editors, Biophysics and Cybernetic Systems, Spartan Books, Washington, D.C., 196[illegible]

Maxon, R.O., Automation in program management (abstr.), ACSM-ASP 1966 Convention Prog., p. 30, 1966

May, J.M., editor, Studies in Disease Ecology, Hafner, New York, 1961, 623 pp.

McBroom, P., Machines cannot think, Science News, v. 90 : 6, 1966

McCarthy, J., Information, Scientific American, v. 215, no. 3 : 64-73, 1966

McCormick, B.H., et al, ILLIAC III -- A processor of visual information, Rept. 183, Dept. of Computer Science, Univ. Ill., Urbana, 1965, 8 pp.

_____ _____, and Richardson, A.M., Design concepts for an information resource center with option of an attached automated laboratory, Rept. 203, Dept. of Computer Science, Univ. Ill., Urbana, 1966, 79 pp.

McCracken, D.D., A Guide to FORTRAN Programming, John Wiley, New York, 1961, 87 pp.

_____, A Guide to ALGOL Programming, John Wiley, New York, 1962, 106 pp.

_____, A Guide to COBOL Programming, John Wiley, New York, 1963, 182 pp.

_____, A Guide to FORTRAN IV Programming, John Wiley, New York, 1965, 151 pp.

_____, et al, Glossary of computer terms, (reprint from Programming, Business Computers), John Wiley, New York, 1964, 24 pp.

_____, and Dorn, W.S., Numerical Methods and FORTRAN Programming, John Wiley, New York, 1964, 457 pp.

McCue, G.A., and Dupree, B.J., Improved FORTRAN IV function contouring program, SID-65-672, Space and Information Systems Div., North American Aviation Inc., 1965, 31 pp.

McElroy, M.N., and Kaesler, R.L., Application of factor analysis to the Upper Cambrian Reagan Sandstone of central and northwest Kansas, The Compass, v. 42 : 168-201, 1965

McGlashan, N.D., The medical geographers work, International Pathology, v. 7, no. 3 : 81-83, 1966

McIntyre, D.B., Trend-surface analysis of noisy data, Kan. Geol. Surv., Computer Contrib. 12, pp. 45-56, 1967

McLean, J.D., Code book for strata data, McLean Paleontological Laboratory, Alexandria, Va., 1962

_____, An application of electronic data processing techniques to paleontology and stratigraphy, McLean Paleontological Laboratory, Alexandria, Va., 1965, 10 pp.

_____, Cumulative index to card catalogs of Foraminifera and Ostracoda, McLean Paleontological Laboratory, Alexandria, Va., 1965, approx. 50 pp.

_____, Formats and procedures for use in data processing systems of the McLean Paleontological Laboratory, McLean Paleontological Laboratory, Alexandria, Va., 1965, 7 pp.

_____, Revision and new procedures for the stratigraphy and geology files, McLean Paleontological Laboratory, Alexandria, Va., 1965, approx. 10 pp.

_____, Resumé of data for Marine Biology Committee of the Marine technology Soci_ _, McLean Paleontological Laboratory, Alexandria, Va., 1965, 12 pp.

Medical Tribune, Computer is aid in all research at University of Utah, Medical Tribune and Medical News, v. 7, no. 54 : 20, 1966

Mendelsohn, M.L., et al, Morphological analysis of cells and chromosomes by digital computer, Proc. 6th IBM Med. Symp., pp. 409-416, 1964

Merriam, D.F., Geology and the computer, New Scientist, (20 May 1965 issue), pp. 513-516, 1965

_____, editor, Computer applications in the earth sciences -- Colloquium on classification procedures, Kan. Geol. Surv., Comp. Contr. 7, 1966, 79 pp.

_____, Geologic use of the computer, Symp. on Recently Developed Geologic Principles and Sedimentation of the Permo-Pennsylvanian of the Rocky Mountains, 20th Annual Conf., Wyo. Geol. Assoc., pp. 109-112, 1966

_____, Computer aids exploration geologists, Oil and Gas J., (23 January 1967 issue), 4 pp., 1967

_____, and Cocke, N.C., editors, Computer applications in the earth sciences -- Colloquium on trend analysis, Kan. Geol. Surv., Comp. Contr. 12, 1967, 62 pp.

_____, and Lippert, R.H., Pattern recognition studies of geology structure using trend-surface analysis, Quart. Colo. Sch. Mines, v. 59 : 237-245, 1964

_____, and _____, Geologic model studies using trend-surface analysis, J. Geol., v. 74 : 344-357, 1966

_____, and Sneath, P.H.A., Quantitative comparison of contour maps, J. Geophys. Res., v. 71 : 1105-1115, 1966

_____, and _____, Comparison of cyclic rock sequences using cross-association, Spec. Publ. 2, Dept. of Geology, Univ. of Kansas, pp. 523-538, 1967

Merritt, C.A., Serving the needs of the information retrieval user, Spring Joint Computer Conference, AFIPS Conf. Proc., v. 30 . 429-432, 1967

Miesch, A.T., Methods of computation for estimating geochemical abundance, U.S. Geol. Surv., Prof. Pap. 574-B, pp. B1-B15, 1967

_____, Theory of error in geochemical data, U.S. Geol. Surv., Prof. Pap. 574-A, pp. A1-A17, 1967

_____, et al, Investigation of geochemical sampling problems by computer simulation, Quart. Colo. Sch. Mines, v. 59 : 131-148, 1964

_____, and Connor, J.J., Investigation of sampling-error effects in geochemical prospecting, U.S. Geol. Surv., Prof. Pap. 475-D, pp. D84-D88, 1964

_____, and Eicher, R.N., A system of statistical computer programs for geologic research, Quart. Colo. Sch. Mines, v. 59 : 259-286, 1964

Miller, A.E., Data transmission -- The total systems concept, PR 7500-022, Auerbach Corp., (reprinted from Data Systems Design), 1964, 4 pp.

Miller, G.B., Production and quality control in map printing at the U.S. Geological Survey (abstr.) ACSM-ASP 1966 Convention Prog., p. 33, 1966

Miller, R.L., and Kahn, J.S., Statistical analysis in the geological sciences, John Wiley, New York, 1962, 483 pp.

Minsky, M.L., Artificial intelligence, Scientific American, v. 215, no. 3 : 246-260, 1966

Monmonier, M.S., The production of shaded maps on the digital computer, Professional Geographer, v. 17, no. 5 : 13-14, 1965

Monroe Internatl. Inc., A brief on FORTRAN XI, Publ. MO-402, Monroe Internatl. Inc., Orange, N. J., 1965, 4 pp.

_____, A brief on QUIKOMP, Publ. MO-401, Monroe Internatl. Inc., Orange, N. J., 1965, 4 pp.

Montgomery, C.J., Computer permits simplified field surveying methods (abstr.), ACSM-ASP 1966 Convention Prog., p. 33, 1966

Moore, G.P., Statistical analysis and functional interpretation of neuronal spike data, Ann. Rev. Physiol., v. 28 : 493-522, 1966

Moser, F., A computer oriented system in stratigraphic analysis, Inst. Science and Technology, Univ. Mich., 1963

Mullins, L.S., Sources of information on medical geography, (reprint), pp. 230-242, 1967

Narasimhan, R., and Fornango, J.P., Some further experiments in the parallel processing of pictures, File No. 616, Digital Computer Lab., Univ. Ill., Urbana, 1964, 13 pp.

National Academy of Sciences, Tropical Health, Publ. 996, Div. of Med. Sci., Natl. Acad. Sci.-Natl. Res. Counc., Washington, D.C., 1962, 540 pp.

National Oceanographic Data Center, Computer programs in oceanography, Publ. C-5, Natl. Oceanogr. Data Ctr., Washington, D.C., 1964, 58 pp.

_____, Instructions for coding and keypunching the geological information form for core, grab, and dredge samples, Publ. M-5 (prov.), Natl. Oceanogr. Data Ctr., Washington, D.C., 1964, 41 pp.

_____, Processing physical and chemical data from oceanographic stations -- Part 1, coding and keypunching, rev. ed., Publ. M-2, Natl. Oceanogr. Data Ctr., Washington, D.C., 1964

_____, Manual for coding and keypunching biological data, Publ. M-4 (prov.), Natl. Oceanogr. Data Ctr., Washington, D.C., 1965

Navy Publications and Printing Service, Electronography, U.S.Navy Publ. and Prntg. Serv., 1964, 8 pp.

Nelson, B., Machine translation -- Committee skeptical over research support, Science, v. 155 : 58-59, 1967

Nelson, D.B., Pick, R.A., and Andrews, K.B., GIM-1, A generalized information management language and computer system, Spring Joint Computer Conference, AFIPS Conf. Proc., v. 30 : 169-173, 1967

Newman, S., Storage and retrieval of contents of technical literature -- Nonchemical information, Res. & Dev. Rept. 4, U.S. Patent Office, Washington, D.C., 1957, 15 pp.

Nooney, G.C., Mathematical models, reality and results, Proc. 5th IBM Med. Symp., pp. 225-242, 1963

Nordbeck, S., Location of Areal Data for Computer Processing, Gleerup, Lund, Sweden, 1962

_____, and Bengtsson, B., Construction of isarithms and isarithmic maps by computers, Nordisk Tidskrift for Informationsbehandling, v. 2, 1964

Bibliography

O'Connor, J., Information retrieval by UNIVAC and by UNIVAC-produced non-mechanized system, part 1, Tech. Rept. 18, Contract Nonr-2297(00), UNIVAC Div., Sperry Rand Corp., Philadelphia, 1957, 98 pp.

Odoroff, M.E., and Abbe, L.M., Use of general hospitals -- demographic and ecologic factors: Public Health Repts., v. 72 : pp. 397-403, 1957

Oettinger, A.G., The uses of computers in science, Scientific American, v. 215, no. 3 : 160-172, 1966

Oliver, L.H., et al, An investigation of the basic processes involved in the manual indexing of scientific documents, Rept., Contract NSF C-422, General Electric Co., Bethesda, Md., 1966, approx. 130 pp.

Ommaya, A.K., et al, A system of coding medical data for punched-card machine retrieval as applied to epilepsy, Epilepsia, v. 5 : 192-200, 1964

Overhage, C.F.J., and Harman, R.J., INTREX -- Report of a Planning Conference on Information Transfer Experiments, MIT Press, Mass. Inst. of Tech., Cambridge, Mass., 1965

Panel on Information Science Technology, First Report of Panel 2 -- Information Sciences Technology, Committee on Scientific and Technical Information, Federal Council for Science and Technology, Working Paper, U.S.Office of Science and Technology, Washington, D.C., 1965, 9 pp.

Parkins, P.V., BioSciences Information Service of Biological Abstracts -- Abstracting and indexing provide input for a dynamic, computer-based information system, (Preprint of draft), 1966

Parks, G.A., editor, Computer in the mineral industries, Stanford Univ. Press, Stanford, Calif., 1964

Patrick, R.L., Wordbook of computer programming terms, Planning Research Corp., Los Angeles, Calif., 1964, 65 pp.

Patterson, G.W., et al, What is a code?, Final Rept. for 1 May 58 - 30 Jun 59 on Project ADAR Task F, Contract DA 36-039-sc-75047, Moore School of Electrical Engineering, Univ. Penn., 1959, 43 pp.

Paulus, J.E., High-speed system for handling of microform documents, Mosler Safe Company, Hamilton, Ohio, 1966, 7 pp.

Pearn, W.C., Finding the ideal cyclothem, Kan. Geol. Surv., Bull. 169, v. 2 : 399-413, 1964

Perkel, D.H., A digital-computer model of nerve-cell functioning, Memorandum RM-4132-NIH, Rand Corp., Santa Monica, Calif., 1964

_____, Applications of a digital computer simulation of a neural network, pp. 37-51 of Mayfield, M., et al, editors, Biophysics & Cybernetic Systems, Spartan Books, Washington, D.C., 1965

_____, Statistical techniques for detecting and classifying neuronal interactions, Proc. of Symp. on Information Processing in Sight Sensory Systems, 1965

_____, et al, Pacemaker neurons -- Effects of regularly spaced synaptic input, Science, v. 145 : 61-63, 1964

_____, and Moore, G.P., A defense of neural modelling, Publ. P-3057, Rand Corp., Santa Monica, Calif., 1965, 9 pp.

Perring, F.H., and Walters, S.M., editors, Atals of the British Flora, Botanical Soc. of British Isles, publ. by Thomas Nelson & Sons Ltd., London, 1962

Perry, K.E. and Aho, E.J., The Calliscope -- A versatile alphanumeric display, Tech. Rept. 212, Contract AF 19(604)5200, Lincoln Lab., Mass. Inst. of Tech., 1959, 18 pp.

Peters, B., Security considerations in a multiprogrammed computer system, Spring Joint Computer Conference, AFIPS Conf. Proc., v. 30 : 283-286, 1967

Petersen, H.W., and Turn, R., System implications of information privacy, Spring Joint Computer Conference, AFIPS Conf. Proc., v. 30 : 291-300, 1967

Phillips, W., et al, Person-matching by electronic methods, Communications of the ACM, v. 5 : 404-407, 1962

Pierce, J.R., The transmission of computer data, Scientific American, v. 215, no. 3 : 144-156, 1966

Pierce, J.W., and Good, D.I., FORTRAN II Program for standard-size analysis of unconsolidated sediments using an IBM 1620 computer, Kan. Geol. Surv., Spec. Dist. Publ. 28, 1966, 19 pp.

Pitts, F.R., Chorology revisited -- Computerwise, Professional Geographer, v. 14, no. 6 : pp. 8-12, 1962

Preston, F.W., et al, The use of statistical communication theory for characterization of porous media, Computers and Operation Research in Mineral Industries, 6th Ann. Symp., Penn. State Univ., 1966, 20 pp.

_____, and Henderson, J.H., Fourier series characterization of cyclic sediments for stratigraphic correlation, Kan. Geol. Surv., Bull. 169, v. 2 : 415-425, 1964

Pritchard, T.C., Automating the engineering product, Graphic Science, v. 8, no. 3 : 21-24, 1966

Project LEX, DOD Manual for building a technical thesaurus, ONR-25, Office of Naval Research, Washington, D.C., 1966, 24 pp.

Public Health Reports, Electronic data processing to detect hospital epidemic, Public Health Repts., v. 82 : 217-218, 1967

Raisz, E., General Cartography, 2nd ed., McGraw-Hill, New York, 1948

Ratynski, M.F., The Air Force computer program acquisition concept, Spring Joint Computer Conference, AFIPS Conf. Proc., v. 30 : 33-44, 1967

Raup, D.M., Computer as aid in describing form in gastropod shells, Science, v. 138 : 150, 1962

_____, and Michelson, A., Theoretical morphology of the coiled shell, Science, v. 147 : 1294-1295, 1965

Read, W.A., Trend-surface analysis of stratigraphic thickness data from some Namurian rocks east of Stirling, Scotland, Scottish J. of Geology, v. 2 : 96-100, 1966

Reitman, W.R., Information-processing models in psychology, Science, v. 144 : 1192-1198, 1964

Rens, F.J., FORTRAN program for coordinate mapping using IBM 7090 computer, Tech. Rept. 10, ONR Task 389-135, Contract Nonr-1228(26), Dept. of Geography, Univ. Mich., 1965, approx. 20 pp.

Rentmeester, L.F., Cybernetics and cartography (abstr.), ACSM-ASP 1966 Convention Progr., p. 39, 1968

Revusky, S.H., Some statistical treatments compatible with individual organism methodology, Rept. 716, U.S. Army Medical Research Lab., Ft. Knox, Ky., 1967, 22 pp.

Reza, F.M., An Introduction to Information Theory, McGraw-Hill, New York, 1961, 496 pp.

Rich, W.H., and Terry, M.S., The industrial control chart applied to the study of epidemics, Public Health Rept., v. 63 : 1501-1511, 1948

Ringertz, N., Possible interrelationships between bovine and human leukemia -- A geographic study, International Pathology, v. 8, no. 2 : 30-31, 1967

Roberts, J.A., The topographic map in a world of computers, The Professional Geographer, v. 14, no. 6 : 12-13, 1962

Rogers, D.J., and Tanimoto, T.T., A computer program for classifying plants, Science, v. 132 : 1115-1118, 1960

Rogers, F.A., 1961, The use of IBM tabulating methods in the analysis of medical data, Minnesota Medicine, v. 44 : 382-386, 1961

Rosen, C.A., Pattern classification by adaptive machines, Science, v. 156 : 38-44, 1967

Rosenfeld, A., and Pfaltz, J.L., Sequential operations in digital picture processing, Journal of the ACM, v. 13 : 471-494, 1966

Ross, D., The mapmakers, Datamation, v. 13, no. 5 : 11, 1967

Sabins, F.F., Computer flow diagram in facies analysis, Amer. Assoc. Petrol. Geol. Bull., v. 47 : 2045-2047, 1963

Sackin, M.J., et al, ALGOL Program for cross-association of nonnumeric sequences using a medium-size computer, Kan. Geol. Surv., Spec. Dist. Publ. 23, 1965, 36 pp.

Sampson, R.J., and Davis, J.C., FORTRAN II trend-surface program with unrestricted input for IBM 1620 computer, Kan. Geol. Surv., Spec. Dist. Publ. 26, 1966, 13 pp.

_____, and _____, Three-dimensional response surface program in FORTRAN II for the IBM 1620 computer, Kan. Geol. Surv., Comp. Contr. 10, 1967, 20 pp.

Sayer, J.S., The use of information technology in research and development planning, PR 7500-055, Auerbach Corp., Philadelphia, 1964

_____, Usage of emerging technology in program execution, PR 7579, Auerbach Corp., Philadelphia, 1964, 22 pp.

Scheele, M., Punch-card methods in research and documentation with special reference to biology, Lib. Sci. and Doc. no. 2, Interscience Publishers, New York, 1962, 282 pp.

Schlager, C.W., Growing importance of map substitutes (abstr.), ACSM-ASP 1966 Convention Prog., p. 32, 1966

Schmid, C.F., and MacCannell, E.H., Basic problems, techniques, and theory of isopleth mapping, J. of Amer. Statistical Assoc., v. 50, 1955

Schmitt, O.H., and Caceres, C.A., editors, Electronic and Computer-Assisted Studies of Biomedical Problems, C.C. Thomas, Springfield, Ill., 1964

Schoenfeld, R.L., The role of a digital computer as a biological instrument, Ann. N.Y. Acad. Sci., v. 115 : 915-942, 1964

Schultz, C.K., et al, Optimization and standardization of information retrieval language and systems, Tech. Status Rept. 1, Contract AF 49(638)835, Univac Div., Sperry Rand Corp., Philadelphia, 1961, 50 pp.

Science News, Machine translation, Science News, v. 91 : 265, 1967

_____, Three-D plotter, Science News, v. 92 : 118, 1967

Shafritz, A.B., and Rose, K., Overflow storage in a store-and-forward digital storage system, Data Systems Engineering, v. 19, no. 1, 1964

Siekmeier, D., Apparatus for the real-time transmission of handwriting and map information to remote displays, Rept. 2900-300-R, Contract DA 36-039-sc-78801, Inst. of Sci. and Tech., Univ. Mich., 1962, 29 pp.

Simpson, M.H., Cataloguing and retrieval of environmental information -- A statement of the problem, AFA R1713, Army Frankford Arsenal, Philadelphia, 1964, 32 pp.

Sisson, R.L., Computer output and display devices, Ann. N.Y. Acad. Sci., v. 115 : 627-643, 1964

Skinner, F.D., Computer graphics -- Where are we? Datamation, v. 12, no. 5, pp. 28-31, 1966

Smillie, K.W., Electronic digital computers and their use in entomology, Entomol. Soc. Canada, Mem. 32, pp. 11-15, 1963

Smith, W.A., Nature and detection of errors in production data collection, Spring Joint Computer Conference, AFIPS Conf. Proc., v. 30 : 425-428, 1967

Snipes, D.S., and Butler, J.R., Digital computer program for identification of minerals by X-ray diffraction, J. Elisha Mitchell Sci. Soc., v. 78 : 97, 1962

Sokal, R.R., and Sneath, P.H.A., Principles of Numerical Taxonomy, W.H. Freeman, San Francisco, 1963, 359 pp.

Soper, J.H., Mapping the distribution of plants by machine, Canadian J. of Botany, v 42 : 1087-1100, 1964

Spacelabs Inc., X-Y data display system, NASA Contractor Rept. CR-460, Contract NAS 4-589, Spacelabs Inc., Van Nuys, Calif., 1966

Speakman, E.D., Automated production planning and control (abstr.), ACSM-ASP 1966 Convention Prog., p. 36, 1966

Spitz, O.T., Generation of orthogonal polynomials for trend surfacing with a digital computer, Computers and Operation Research in Mineral Industries, 6th Ann. Symposium, Penn. State Univ., 1966, 6 pp.

Stacy, R.W., and Waxman, B.D., Computers in Biomedical Research, Academic Press, New York, vols. 1 and 2, 1965

Stamp, L.D., The Geography of Life and Death, Fontana Library, Collins, London, 1964

Statland, N., Methods of evaluating computer systems performance, Computers and Automation, v. 13, no. 2, 1964

_____, and Hillegass, J.R., A survey of computer input-output equipment, PR 7533, Auerbach Corp., (repr. from Data Processing Yearbook), 1963(?), 20 pp.

_____, _____, and _____, Random access storage devices, PR 7556, Auerbach Corp., (reprint from Datamation), 1963, 9 pp.

Steakley, J.E., Automated color-separation system (abstr.), ACSM-ASP 1966 Convention Prog., p. 46, 1966

Steinberg, A., and Paine, L.W., Methods and techniques of data conversion, Ann. N.Y. Acad. Sci., v. 115 : 614-626, 1964

Sterling, T., and Pollack, S., MEDCOMP handbook of computer applications in biology and medicine -- Part 1, Statistical systems, Medical Computing Center, College of Medicine, Univ. Cincinnati, 196[illegible], 256 p.

Strachey, C., System analysis and programming, Scientific American, v. 215, no. 3 : 112-124, 1966

Sublette, I.E., Recognition of class membership by means of weak, statistically dependent features, AMRL-TR-66-174, U.S. Air Force Aerospace Medical Research Lab., Wright-Patterson A.F.B., Ohio, 1966, 39 pp.

Suppes, P., The uses of computers in education, Scientific American, v. 215, no. 3 : 206-220, 1966

Sutherland, I.E., Computer graphics, Datamation, v. 12, no. 5 : 22-27 196[illegible]

_____, Computer inputs and outputs, Scientific American, v. 215, no. 3 : 86-96, 1966

Switzer, P., et al, Statistical analysis of ocean terrain and contour plotting procedures, A.D. Little, Cambridge, 1964

Tatch, D., Automatic encoding of medical diagnoses, Proc. 6th IBM Med. Symp., pp. 545-551, 1964

Taube, M., Computers and common sense -- The myth of thinking machines, Columbia Univ. Press, New York, 1961

Tewinkel, G.C., Block analytic aerotriangulation (abstr.), ACSM-ASP 1966 Convention Prog., p. 38, 1966

Thoma, J. A., Simple and rapid method for the coding of punched cards, Science, v. 137 : 278-279, 1962

Thompson, E.T., and Hayden, A.C., Standard Nomenclature of Diseases and Operations, 5th ed., McGraw-Hill, New York, 1961, 964 pp.

Tjalma, R.A., et al, Clinical records systems and data retrieval function in veterinary medicine -- A proposal for systematic data programming, J. Amer. Vet. Med. Assn., v. 145 : 1189-1197, 1964

Tobler, W.R., Automation and cartography, Geographical Rev., v. 49 : 526-534, 1959

_____, Geographical ordering of information, The Canadian Geographer, v. 7, no. 4 : 203-205, 1963

_____, Automation in the preparation of thematic maps, The Cartographic J. (reprint, 7 pp.), 1964

_____, Computation of the correspondence of geographical patterns, Papers of the Regional Science Association, pp. 131-139, 1965(?)

_____, Numerical map generalization, Univ. Mich., Department of Geography, Michigan Inter-University Community of Mathematical Geographers Discussion Paper 8, Part 1, 1966, 25 pp.

_____, Spectral analysis of spatial series (reprint, 1967, 8 pp.

Tolles, W.E., editor, Computers in medicine and biology, Ann. N.Y. Acad. Sci., v. 115 : 543-1140, 1964

Toomey, D.F., Application of factor analysis to a facies study of the Leavenworth Limestone (Pennsylvanian-Virgilian) of Kansas and environs, Kan. Geol. Surv., Spec. Dist. Publ. 27, 1966, 28 pp.

Turner, A.H., editor, Computers in medicine bibliography, Dept. of Radiology and Medical Center Library, School of Medicine, Univ. Miss., 1965, 164 pp.

_____, and Schmidt, D.A., Computers in medicine bibliography, rev. ed., Dept. of Radiology and Medical Center Library, Univ. Miss., 1966

Univac Division, Electronic data-processing for the line official, Publ. U 2448E, Univac Div., Sperry Rand Corp., Philadelphia, 1960, 86 pp.

_____, Mighty new servant to the mind of man, Publ. U-4323, Univac Div., Sperry Rand Corp., Philadelphia, 1964, 30 pp.

University of Pittsburgh Department of Geography, Arthropod distribution maps, sponsored by U.S. Army Natick Laboratories, Natick, Mass., 1965-1967

Urban Renewal Service, Using computer graphics in community renewal -- Computer methods of graphing, data positioning and symbolic mapping, Community Renewal Program Guide 1, Urban Renewal Administration, Washington, D.C., 1963, approx. 200 pp.

U.S. Geological Survey, National Atlas, (in preparation)

Vallbona, C., et al, System for processing clinical research data, Proc. 6th IBM Med. Symp., pp. 437-486, 1964

Vitro Laboratories, Plan of action for U.S. Naval Oceanographic Office Library study, Rept., Contract N62306-1828, Vitro Corp. of America, Silver Spring, Md., 1966, approx. 20 pp.

_____, System performance specification for U.S. Naval Oceanographic Office Library study, Rept., Contract N62306-1828, Vitro Corp. of America, Silver Spring, Md., 1966, approx. 20 pp.

_____, The user requirements specification for U.S. Naval Oceanographic Office Library study, Rept., Contract N62306-1828, Vitro Corp. of America, Silver Spring, Md., 1966, approx. 20 pp.

Vogel, P., An inventory of geographic research of the humid tropic environment, Contract DA49-092-ARO-33, Texas Instruments Inc., Dallas, Tex., 1965, 518 pp.

Walker, A., editor, Proceedings of the Symposium on Development and Management of a Computer-Centered Data Base, System Development Corp., Santa Monica, Calif., 1964, 133 pp.

Walker, A.R.P., Complexity of nutritional problems in developing countries, International Pathology, v. 8, no. 4 : 71-73, 1967

Wall, E., The distribution of term usage in manipulative indexes, American Documentation, v. 15, no. 2, 1964

Wallace, R.E., Available communications equipment and status of the art, PR 7685, Auerbach Corp., Philadelphia, 1964, 11 pp.

Warden, J., CalComp Plotter Manual, Informal Manuscript Rept. Misc.-1-65, U.S. Naval Oceanographic Office, Washington, D.C., 1965, 45 pp.

Ware, W.H., Security and privacy in computer systems, Spring Joint Computer Conference, AFIPS Conf. Proc. v. 30 : 279-282, 1967

_____, Security and privacy -- Similarities and differences, Spring Joint Computer Conference, AFIPS Conf. Proc., v. 30 : 287-290, 1967

Warren, H.V., Medical geology and geography, Science, v. 148 : 534-539, 1965

Watson, C., Computer generation of word association maps for man-machine communication, Publ. SP-1153, System Development Corp., Santa Monica, Calif., 1963, 24 pp.

Watson, D.E., et al, Compilation of magnetic charts by analytical procedures utilizing computer techniques (abstr.), ACSM-ASP 1966 Convention Prog., p. 50, 1966

Watson, F.R., Coordination of data and compilation of forms for the computer, in Proc. Conf. Consultants, Guests, & Resident Staff, 19-20 September 1963, Center for Zoonoses Research, Univ. Ill., Urbana, 1963

Webb, G.N., Communicating biological information to the 1401 computer, coding, editing, and interfacing -- Problems and results, Proc. 6th IBM Med. Symp., pp. 69-80, 1964

Wegmueller, F., Codeless scanning -- A new method of automatic documentation [Germ.] : Experientia, v. 1 : 383-384, 1960

Weik, M.H., and Confer, V.J., Survey of scientific and technical information retrieval schemes within the Department of the Army, BRL Rept. 1169, Aberdeen Proving Ground, Md., 1962, 95 pp.

Werling, R., Action-oriented information systems, Datamation, v. 13, no. 6 : 57-65, 1967

Whiteman, I.R., The role of computers in handling aerospace systems human factors task data, Rept. AMRL-TR 65-206, Computer Concepts Inc., 1965, 182 pp.

Bibliography

Whitfield, H., Application of a toxonomy computer programme to disease classification (abstr.), Biometrics, v. 19 : 368, 1963

Whitlock, L.S., Information coding and retrieval of nematology literature on IBM 1620 computer (abstr.), Dissert. Abst., v. 24 : 927, 1963

Wolf, M., Computational techniques in linguistic geography, (reprint), 1966

Woodbury, M.A., Time series factor analysis, Proc. 2nd IBM Med. Symp., pp. 385-390, 1960

World Health Organization, Trends in the study of morbidity and mortality, Public Health Paper 27, 1965, 196 pp.

_____, Computers in medicine, W.H.O. Chronicle, v. 21 : 100-111, 1967

Wright, H.T., Marienfeld, C.J., and Silberg, S.L., "Place" in environmental epidemiology of rectangular coordinate method, Public Health Reports, v. 83, no. 5 : 427-434, May 1968

Wright, J.K., A proposed Atlas of Diseases, Appendix 1, Cartographic considerations, Gecgr. Rev., v. 34, 1944

Yamada, S., and Fornango, J.P., Experimental results for local filtering of digitized pictures, Rept. 184, Dept. of Computer Science, Univ. Ill., Urbana, 1965, 44 pp.

Yoder, F.D., Data processing in public health, Proc. 4th IBM Med. Symp., pp. 193-204, 1962

Yoder, R.D., Tulane Information Processing System, version 1, Monogr. 1, Computer Science Series, Tulane Univ., 1965

Zimmer, H., Preparing psycholphysiologic analog information for the digital computer, Behav. Sci., v. 6 : 161-164, 1961

Zubryn, E., Electronic herbarium, Science News, v. 92 : 161, 1967

A recent publication which became available too late to include in its proper (alphabetical) place, but that is too important to omit from this listing is:

Lindberg, D.A.B., The Computer and Medical Care, Charles C. Thomas, Springfield, Ill., 1968, 210 pp.

Appendix

The Appendix includes the following.

Glossary

-- *Computer processing terms* A-2

-- *Biomedical terms* A-6

Data sources

-- *Narrative and tabular* A-10

-- *Published maps* A-16

Schistosomiasis: A-25

Leptospirosis: A-28

Glossary

This monograph incorporates technical information derived from several disciplines, each with its own jargon. In the interests of effective communication, some of these terms which have "special" meaning have been selected for brief explanation here. We realize full well that epidemiologic terms do not need to be explained to the epidemiologist, nor cartographic ones to the cartographer -- but the cartographer may find a definition of certain epidemiologic terms helpful, and vice versa. Then, too, some of the terms listed here have varied meanings, even within the primary discipline, depending upon who uses them and in what context. We have tried to be precise and consistent in our usage of these terms, adhering to the meanings given here.

For convenience, the glossary is divided into two parts: Part one considers data-processing terms; Part two, biomedical terms.

MOD DATA PROCESSING TERMS

Block Diagram -- A representation of spatial or areal relationships on the earth's surface that is drawn obliquely to that surface but which, otherwise, is the same as a map. When used to present geologic data, block diagrams usually show a horizontal surface area and two vertical cross-sections, but when used to present disease data, the two vertical cross-sections, often unnecessary, are often omitted.

CEN -- see Computer Evaluation Number.

C-MOF -- see Common-MOF.

Common-MOF (C-MOF) -- A MOF which should, and usually does, accompany (as a necessary descriptive element), or should be common to, every data point or bit of mappable data. In the MOD system only six C-MOF's are recognized (see p.

Computer Evaluation Number (CEN) -- A number calculated by the MOD computer system, according to an appropriate algorithm, to indicate the relative reliability of each data point that is input to the system.

Data Point -- A specific geographic locality where a particular factor/aspect/facet of the total disease/environmental situation has been determined/observed/measured, and the result/evaluation/value expressed in some qualitative/quantitative forms. In manual mapping procedures a data point is represented by a dot at the location of the point, with a symbol beside or overprinted on the dot to indicate the value of the data point. In the MOD system a data point consists of a geographic location (LOC), a data-point value (VAL), a factor (HOF or POF), and narrative (NAR).

Disease Map -- A map showing some aspect, facet, or factor of the total disease situation (ecology).

Factor -- Alphabetic and/or numeric symbols naming/describing exactly what part/aspect/facet of the total disease/environmental situation is being evaluated (i.e., given a VAL) at (the LOC of) the specific data point. Factor is a general term that includes LOF's, MOF's, HOF's, and POF's, and is one of the three essential components of a data point.

Graph -- Straight/curved lines, points, and words/numbers, all representing numerical data which express the relationship among specific variables.

High-Order Factor (HOF) -- A specific combination of LOF's in which each LOF belongs to (is drawn from) a different MOF; i.e., a specific combination of LOF's to which no MOF contributes more than one LOF (see p. 4-7).

HOF -- see High-Order Factor.

LA -- see Latitude.

Latitude (LA) -- Angular distance along earth's surface as measured north or south from equator.

LO -- see Longitude.

LOC -- see Location (Geographic).

Location (Geographic) (LOC) -- The exact geographic position, stated as precisely as possible, of the data point. Location is one of the three essential components of a data point (i.e., each bit of mappable data).

LOF -- see Low-Order Factor.

Longitude (LO) -- Angular distance along earth's surface as measured east or west from Greenwich meridian.

Low-Order Factor (LOF) -- The most specific possible name or description or a particular disease/environmental situation.

Map -- A graphic/visual presentation, on a geographic-coordinate basis, of the information imparted by a particular set of specific data points. A map is a representation of spatial or areal relationships on the earth's surface, drawn perpendicularly to that surface and according to a rigorous grid pattern and scale so that there results no nonsystematic distortion of size, shape, distance, and neighbors. A map is, essentially, a three-variable graph in which X = LO, Y = LA, and Z = value of whatever factor is being mapped.

Middle-Order Factor (MOF) -- The set of all LOF's which describe the same aspect/facet of disease/environmental situations. (See p.

MOF -- see Middle-Order Factor.

Multi-LOF MOF -- A MOF which can contain more than one LOF for each data point. For example, the MOF "Specific Disease Agent", can include several LOF's: "Leptospira pomona, L. canicola, and L. seiroe",all at one data point.

NAR -- see Narrative.

Narrative (NAR) -- Supporting, nonmappable prose/narrative/textual information or data associated with a specific data point.

O-MOF -- see Optional-MOF.

Optional-MOF (O-MOF) -- A MOF which need not fit into every possible disease/environmental data point and which, in a sense then, is optional. This category includes all MOF's except Common-MOF's.

POF -- see Poly-Order Factor.

Poly-Order Factor (POF) -- A specific combination of LOF's in which at least two LOF's belong to (are drawn from) the same MOF; i.e., a specific combination of LOF's, to which at least one MOF contributes more than one LOF. (See p.

Primary Data Point -- A data point extracted from text that originally reported that data, i.e., from its primary source document.

Qualitative -- Expressed or denoted by alphabetic symbols or words. When so expressed, LOF's (and MOF's containing them) and VAL's can be termed "qualitative".

Quantitative -- Expressed or denoted by numeric symbols or numerals. When so expressed, LOF's (and MOF's containing them) and VAL's can be termed "quantitative".

Report (Hard-Copy Report) -- Printed words and/or numbers arranged in listings, tables, or narrative-like prose.

Secondary Data Point -- A data point extracted from text, i.e., from its secondary source document referencing a previous report of that data.

Single-LOF MOF -- A MOF which can contain only one LOF for each data point. For example, the MOF, "Total Annual Rainfall", can include only one LOF, e.g., "13 inches", at one data point.

System -- Used in two senses in this report, but differentiated by context:
1. MOD system, consisting of the personnel, procedures, programs, and equipment (including computer), integrated to perform mapping of disease/environmental data;
2. MOD computer system, consisting only of the various programs and equipment mentioned above.

System Analysis -- Investigation of an activity or procedure to determine what that activity/procedure must accomplish, what it has available to it, and how its necessary operations may best be accomplished (either manually or by computer).

System Design -- Planning of a system by specifying the characteristics, actions, and relationships among the various parts (personnel, programs, and equipment) of the system.

System Implementation -- Actual construction of a system, including production of programs, installation of equipment, hiring and training of personnel which -- all together -- comprise the functional system.

System Operation -- Operation of the system on a regular production basis in which the personnel utilize the system procedures, programs, and equipment to accomplish the task for which the system was designed and implemented.

VAL -- see Value (for Data Point).

Value (for Data Point) (VAL) -- An alphabetic and/or numeric symbol expressing the precise character/condition of that aspect/factor (of

the disease/environmental situation) being considered at (the LOC of) the specific data point. Value is one of the three essential components of a data point.

BIOMEDICAL TERMS

Carriers (of disease) are infected persons who harbor an infectious agent and who are capable of transmitting this agent, but who have no obvious manifestations of the disease.

Contamination: See infection.

Ecology is the study of relationships between living organisms and their habitat -- an analysis of the biodynamics within communities. The ecology of disease is the study of relationships among hosts, disease agents, and their environments.

Endemic/enzootic diseases are those which are present in a given community (human beings/animals) at all times, but at a low level. Hyperendemic/hyperenzootic diseases are those continuously present at a high rate in human beings/animals.

Epidemic/epizootic diseases are those intermittently present at a high rate in a (relatively) small area. They may be diseases new to the community or diseases that were continually or sporadically present at low levels, but that are now occurring at a much higher rate than usual. The suffix -demic, relates to human beings; the suffix -zootic, relates to animals.

Geographic pathology is, in a sense, a kind of comparative pathology -- one in which place (rather than species) is the primary variable. It is concerned with *what* diseases occur *where*, and *why*. It is also concerned with the reasons why the "same" disease (in terms of causative agent) may behave quite differently in different parts of the world.

Host is an animal (or a plant) which harbors an infectious agent. The host may or may not suffer disease as a result.

Hyperendemic/hyperenzootic: See endemic/enzootic.

Immunity may be relative or absolute. Absolute immunity protects against disease; relative immunity attenuates the disease. Immunity is of two

types: innate immunity, relating to those factors *inherent* in the body that act to resist an infectious agent -- and acquired immunity, a state of increased resistance that is related to the presence of specific (acquired) antibodies against the infectious agent. The specific antibodies can come about "naturally", i.e., as a consequence of infection, or they can be produced "artificially", as from vaccination.

Incidence and prevalence are important terms, often misused. They both relate to the time during or at which a disease is studied. Events such as (clinical) onset of the disease, or birth, or death occur at a precise point in time, but various disease states, e.g., leptospirosis, or schistosomiasis, or diabetes, exist over varying periods of time, perhaps years. Incidence describes the number of events (related to the occurrence, i.e., the *onset* of a disease) which took place during a specified time. Prevalence, on the other hand, refers to the number of cases of a particular disease which *existed* -- at any stage -- at (or during) a particular time in a given population. For example: the incidence of leptospirosis in human beings in country X was determined as 127 cases per 100,000 population for 1964. The figure 127 includes (properly) only those cases that *began* during 1964. The prevalence of leptospirosis in human beings in country X was determined as 147 per 100,000 population for the year 1964. (In this hypothetical study, much care was taken not to count the same diseased person more than once.) This figure, 147, includes those cases which had their onset before 1 January 1964, but which persisted into the time period under observation, i.e., 1 January through December 31, 1964. From these particular incidence and prevalence figures, one could infer that leptospirosis was a disease that probably persisted for several weeks, since 27/147 cases observed in one year had their beginning before that year. The more chronic the disease, obviously, the greater the disparity between incidence and prevalence figures. "Point prevalence" refers to the number of cases present during a very short period of observation. (Short-term field surveys usually determine point prevalence, i.e., the number of cases -- at all stages -- present at the time that the particular population was examined. "Period prevalence" refers to the number of cases present during a (longer) specified period of observation.

Infection is a disease state resulting from an (infectious) agent -- virus, bacteria, spirochete, yeast, fungus, or (animal) parasite -- living in the host and producing some sort of defensive reaction by the host. The infection is not always apparent -- either to the patient or his doctor. Sometimes, although there are no signs or symptoms, there is laboratory evidence of the reaction, e.g., the presence of antibodies specific for the infectious agent. In this latter instance the infection is said to be silent, or inapparent or (sometimes) "subclinical". Infection is to be sharply distinguished from contamination, a situation in which infectious agents may be "resting"

on the exterior surfaces of the body or upon articles of clothing, etc. The term contamination also applies to conditions in which the infectious agents are contained within soil or water or food.

Infestation is ordinarily applied to ectoparasites and describes a host-parasite relationship in which the parasite lives on the surface of the host. In some instances, e.g., scabies, the parasite may temporarily invade and inhabit the superficial tissues of the host. (It is also possible for an ectoparasite to live on such "internal" surfaces as the intestinal mucosa, but, again, it must not invade the tissues of the host, otherwise the relationship would be one of infection.)

Morbidity relates to the (non-lethal) manifestations of disease. Morbidity rates are, in essence, "sick rates".

Mortality, in relation to disease, concerns the lethality of the disease. As a rule, mortality rate pertains to the ratio of number of deaths from a given disease to the total population under study. See "Rates".

Pandemic/panzootic diseases are those intermittently present at a high rate over a very large area, e.g., several countries -- or diseases continuously present but now at a much higher rate than usual. In a sense, a pandemic is a very widespread epidemic. As before, the suffix, -demic, relates to human beings, the suffix, -zootic, relates to animals.

Parasite, in its broadest sense, includes all living agents that live in or on a host, deriving benefit from the host, but not necessarily producing disease. These agents include viruses, bacteria, spirochetes, yeasts and fungi, as well as parasitic agents (in the narrow sense). In its restricted meaning, the term parasite refers only to ANIMAL agents; viruses, bacteria, spirochetes, and yeasts and fungi are excluded.

Pathogenicity refers to the capacity of an infectious agent to cause disease in a susceptible host.

Pathology is the study of disease, with particular concern for its cause (*etiology*), the mechanisms of its development (*pathogenesis*), and the nature of its effects, especially those which are of value in establishing specific diagnosis.

Portal of entry refers to the route through which the infectious agent enters the body, e.g., by inhalation, by ingestion, through a traumatic wound, injected in the course of a mosquito bite, etc.

Prevalence: See incidence.

Rates is an expression of the frequency with which a certain event or circumstance occurs in relation to time. There are many kinds of rates: death rates ordinarily relate mortality (often from a specific disease) to the entire population at risk, and are usually expressed as: [deaths (per year) / population] X 1,000; case fatality rates describe the mortality from a specific disease as: [number of deaths from the disease / total number of cases of the disease] X 100.

Reservoir (of infection) is closely related to source, but differs in an important respect. It is a place within a host-parasite system where the population of the infectious agent is *maintained,* and from which a vector commonly transmits it to a susceptible host. (It is not necessary that the infectious agent multiply in a reservoir.)

Source (of infection) is a place, animate or inanimate, where the infectious agent(s) is generated (i.e., *multiplies*), and from where it may be introduced into a new area.

Sporadic cases of disease are those intermittently present, at a low rate.

Vector is an object, either animate or inanimate, that transports an infectious agent to its host. Vectors may be mechanical or biologic. Biologic vectors may also make an essential contribution to the growth and/or development of the parasite, e.g., the mosquito in malaria. (When they make this essential contribution they are called intermediate hosts.)

Virulence is a term somewhat comparable to pathogenicity but it pertains to the ability of the organism to produce severe illness. A highly virulent agent is one which is likely to produce a very serious infection.

Zoonoses are diseases of animals that may be transmitted to man.

> But o'er anxious thought you'll find of no avail,
> For there precisely where ideas fail,
> A word comes opportunity into play
> Most admirable weapons words are found,
> On words a system we securely ground,
> In words we can conveniently believe,
> Not of a single jot can we a word bereave."
>
> Johann Wolfgang von Goethe

Data sources

Section IV has considered data characteristics, and Section V data collection, and we shall not repeat here the things that were discussed there. This brief consideration of data sources -- where to find them and how to get them -- focuses on health/disease data. It makes no attempt to be exhaustive; the particular sources cited are meant to serve simply as examples.

BOOKS are enormously valuable data sources, but often of more historic than current use, and this applies to even the most recent publications (reflecting the time lag between gathering the data, converting it to manuscript, and getting the manuscript published). Nevertheless, books such as Studies in Disease Ecology, edited by Jacques M. May, Hafner, New York, 1961, and Tropical Health -- A Report on a Study of Needs and Resources, Publication 996 of the National Academy of Sciences - National Research Council, Washington, 1962, can be extremely useful.

The vast numbers of PUBLISHED ARTICLES listed in the Cumulative Index Medicus are relatively accessible, but, far too often, the title of the paper does not reflect some of the crucially important data that it contains. The "demand search" function of the MEDLARS (Medical Literature Analyses and Retrieval System -- National Medical Library, Washington) helps to overcome this difficulty, but to only a limited extent. In a recent report (Jan. 1968), Evaluation of the MEDLARS Demands Search Service, it was stated that: "... the system is operating, on the average, at about 58% recall and 50% precision". Furthermore, MEDLARS is primarily concerned with *key words*, not content, per se.

Another approach to the data source selection problem is effectively used by such abstracting services as Chemical Abstracts and Biological Abstracts (BioSciences Information Service), and it is possible to arrange for special services with such organizations as these. Under these conditions, the user can get a good idea of the article's content, and selectivity becomes progressively more precise -- depending upon how much one wishes to pay for this precision.

There is a tremendous amount of important data, the sources of which are not included in Cumulative Index Medicus, i.e., non-indexed data sources, and it is a major problem just being aware that some of these reports exist. The remainder of this general discussion of data sources will concern this category of information.

* * *

There are three principal approaches to the non-indexed data material. One can look for these data by: (1) geographic area, or (2) disease or environmental factors, per se, or (3) a primary data source, e.g., a specific research institute, a specific hospital, a specific individual, etc. Any effort to build a data base, in depth, should use all three of these approaches, albeit they may overlap to considerable extent.

As to GEOGRAPHIC AREA, the local government is an important first source. In many of the so-called developing countries, the principal source of "official" data for the country will be the Ministry of Health, and there may be a series of annual reports that provide valuable current information as well as data that allows important historic perspective. In "developed" countries, such as the United States, there are many many governmental sources of data pertaining to disease-environmental situations of that country: the Bureau of Census, the National Institutes of Health, the Communicable Disease Center, the United States Army, to name but a few. The various State Health Departments have additional, more detailed information and, finally, there may be still more precise data available from specific County Health units.

Important non-governmental sources of (local) data are also numerous: Universities, Research Institutes, Organizations such as the American Medical Association, the American Cancer Society, the New York State Life Insurance Company, etc. etc.

Turning to sources of data which are international in scope, WHO, PAHO, FAO, are very important primary sources for most of the world. Coverage may be very broad (geographically), or quite restricted, e.g., WHO's report, Studies on Immunoglobulins of Nigerians, 1966. Often the reports concentrate on a particular disease or condition, e.g., WHO's Malaria Yearbooks and PAHO's Immunologic Aspects of Parasitic Infection, 1967. In addition to the official international organizations, those individual governments that have had a long interest in international affairs are rich sources of information dealing with other countries. In the United States, for example, much information is available from the National Institutes of Health, the Department of State (especially AID, and the Bureau of Intelligence and Research), the Communicable Disease Center, and the Department of Defense (consider this report, for example). Periodic reports from DOD, or Army, or Navy, or Air Force units -- and other Governmental agencies are valuable data sources and we list below illustrative examples of these:

406th Medical Laboratory Professional Report (annual), United States Army Medical Command, Japan.

Annual Progress Report, SEATO Medical Research Laboratory Clinical Research Center, Bangkok, Thailand.

Annual Work Unit Progress Report from the various Naval Medical Research Units (NAMRU's), e.g., *Serologic, Epidemiologic, and Vaccine Studies on Meningococcal Meningitis* (report of a study carried out in Egypt and Morroco).

Annual Research Project Report, Armed Forces Institute of Pathology (AFIP), Washington.

Annual Reports of the U.S.Army Medical Research Unit's Institute for Medical Research at Kuala Lampur, Malaysia.

Annual Progress Reports, U.S.Army Research Institute of Environmental Medicine, Natick, Mass. (of the U.S. Army Medical Research and Development Command).

National Communicable Disease Center's Morbidity and Mortality Reports and (sporadic) Surveillance Reports.

Forest Service Research Papers, e.g., *Weather in the Luquillo Mountains of Puerto Rico* (250 pages), by C.B.Bresco -- The Institute of Tropical Forestry, Forest Service, U.S.Department of Agriculture

SPECIFIC PRIMARY DATA SOURCES can be very valuable, for example, the Rockefeller Institute, Johns Hopkins University (the Geographic Epidemiology Unit of the School of Hygiene and Public Health), the Institute of Public Health of Iran, the Liverpool School of Tropical Medicine, the Walter Reed Army Institute of Research, the Technical Assistance Information Clearing House, American Council of Voluntary Agencies (44 E. 23rd Stree, New York, N. Y. 10010), etc. Several illustrations of data available from these sources follow:

Annual Report on the Research Activities of the Liberian Institute of the American Foundation for Tropical Medicine.

Tulane University (New Orleans)/Universidad Del Valle (Cali, Colombia) periodic Progress Reports -- an ICMRT (NIH) supported program.

Relation of Geology and Trace Elements to Nutrition (based on papers presented at a Symposium held at the Annual Meeting of the Geological Society of America, New York, 1963), edited by H.L.Cannon and D.F.Davidson.

Proceedings of the 7th International Congress of Tropical Medicine and Malaria (four volumes), Rio de Janeiro, Sept. 1-11, 1963

The Physical Environments and Agriculture of Thailand, by M.Y.Nattonson, a publication of the American Institute of Crop Ecology, Washington, 1963.

Then there are collections of more general data, some of which are reissued periodically, bringing the information up-to-date. For example:

Health Data Publications, pertaining to individual countries (some 44 have been released), prepared by the Department of Health Data, Division of Preventive Medicine, The Walter Reed Army Institute of Research.

Special Warfare Area Handbooks for various countries, prepared by Foreign Areas Studies Division, Special Operations Research Office of the American University, Washington (operating under contract with the Department of the Army).

U.S.Army Area Handbooks for various countries (Department of the Army pamphlets).

The World Food Problem, a two volume report of the President's Science Advisory Committee, 1967.

These many valuable reports, are not "lost" even though they are not indexed in the Quarterly Cumulative Index. There is a variety of other indices available -- if one knows where to find them. For example:

U.S.Government Research and Development Reports, a semi-monthly abstract journal produced by the Clearing House for Federal and Technical Information of the U.S.Department of Commerce.

Air Force Scientific Research Bibliography (abstracts of all USAF Office of Scientific Research Supported Research Projects).

ILSE -- Interagency Life Sciences Supporting Research and Technology Exchange, prepared by Documentation, Inc., (abstracts NASA and DOD Research Work-Units in Life Sciences).

Pesticides Documentation Bulletin, National Agriculture Library, U.S.Department of Agriculture.

But there are many potentially valuable publications that escape the usual indexing mechanisms, and these may be of crucial value in connection with certain geographic areas. For example:

The South African Institute for Medical Research -- Annual Reports

Conibucão Ao Estudo da Patologia Das Arborviruses by Domingos De Paola (91 pages), Rio de Janeiro, 1964.

(A thesis presented in partial fulfillment of requirements to attain Professorial status (docencia-livre), privately published and distributed by the author. Particularly in Latin America, many (medical) doctoral and professorial theses -- such as this one -- represent hidden sources of very valuable local disease-environmental data.

Many SPECIFIC INDIVIDUALS could be mentioned w[illegible]e store of information and, often, even more important, their knowledge of where to find particular information, could be given. As an example, it seems appropriate to mention the single individual who has been most helpful to this project in providing information about leptospirosis: Dr. Aaron D. Alexander (Walter Reed Army Institute of Research). In accumulating information in depth -- and in valuating the depth of information coverage one has achieved -- there is no substitute for wise, skillful, informed consultants who, as a rule, not only help to identify defects in the data base, but give sound advice as to how they can be corrected.

Language can be an important barrier, particularly when the data is reported in a language with which few are familiar, and there is no question but that many valuable data are lost to general use because of this. English translations are available for some of the larger more important reports, for example: Natural Foci of Transmissible Diseases as Related to Territorial Epidemiology of Zooanthroponoses -- USSR, a 254 page English translation prepared under contract for the Joint Publications Research Service, clearing house for Federal Scientific and Technical Information, U.S.Department of Commerce. But such "free" translations service only (partially) covers a rather highly selective subject area.

* * *

The leptospirosis source documents used in this study were located by a combination of general, comprehensive surveillance of epidemiologic

literature and personal contacts with leptospiral workers. In addition to searching standard bibliographies and abstract journals for these documents, LTC Watson's group obtained machine-computer-produced bibliographic lists from such agencies as the National Library of Medicine, the Defense Documentation Center, the Army Research Office, and the Military Entomological Information Service. The World Health Organization and the Pan American Health Organization were also used extensively as data sources.

• • • • • • • • • •

PUBLISHED MAPS

In ordinary usage, MOD generated mapped disease and environmental data would be directly related to (overlaid on) base maps showing topographic features, geologic characteristics, land usage, population density, etc., etc. For this reason such maps represent a very important part of MOD data even though the data contained therein will not (as a rule) be subjected to computer processing.

As a part of our data collection study we have carried out a pilot investigation to determine availability and source of (potential) base maps -- and their usefulness. This study was undertaken with the help of Dr. R. Warwick Armstrong, Assistant Professor of Geography, the University of Ill. Dr. Armstrong has been a valuable consultant medical geographer on the MOD project since its inception, and it was through his cooperation that Mr Gary G. Gullett, graduate student of the Department of Geography, was employed part time to evaluate source material in the form of published maps, floristic atlases, etc., which might serve as base maps or otherwise contribute in an important way to the data base of MOD system

Mr. Gullett made this survey under the guidance of Dr. Armstrong, utilizing the extensive libraries of the University of Illinois, at Urbana libraries, which include one of the largest collections of cartographic/geographic data in the United States. Mr. Gullett's final report, which follows, is presented in its entirety.

In addition (following Mr. Gullett's report) there is listed the major primary sources of maps which were found to be most useful to the MOD project.

Appendix

1. Purpose The purpose of this report is to indicate the location, quality, and quantity of specific maps in the various libraries of the University of Illinois, Urbana, Illinois, which would serve as a data source for the MOD project.

2. Scope The scope of this study includes the searching, assessment, and transmittal of cartographical material concerned with the following categories:

(1) climate

(2) vegetation patterns and associations

(3) ecological studies

(4) soil distributions and associations

(5) entomology

(6) zoology

(7) forestry and forest associations

(8) land use and land use patterns

(9) hydrology

(10) general topography

The searching for, and assessment, and transmittal of the above categories was pursued with reference only to the following geographical areas, and at scales ranging from 1:25,000 to 1:10,000,000:

(1) World (entire)

(2) World (major section, continent, ocean)

(3) Southeast Asia

(4) Thailand

(5) Burma

(6) French Indo-China

. . . . continued

(7) Philippines

(8) Malaya

(9) The Midwest of the U.S.

(10) Illinois (entire)

(11) Southern Illinois (Quadri-County area)

3. Method of Collecting Data Information was obtained exclusively by personal investigation of the cartographical holdings of the several departmental libraries at the University of Illinois. Every appropriate map was checked out of the libraries and the necessary information transcribed onto data-collection forms and, where appropriate, the legends reproduced by xerox. Where there were a series of maps, such as the 1:1,000,000 coverage of the world, only one map in the series was selected, and its representative legend and inherent information transcribed. In addition, a notation was made, marking the number of maps in that series, their extent of coverage, and their call numbers. All necessary information was transmitted to the MOD project headquarters in Washington, D.C. on data-collection forms, and the xeroxed legend attached. A copy of the data-collection form is attached to this report.

4. Results The Map and Geography Library was by far the greatest source of maps for this project. Over 140 different maps were obtained from this library, and many of these 140 "maps" comprised a series of sheets. The total number of applicable sheets was approximately 1,400. The quality of these maps is high and they are relatively up-to-date. They deal, in one way or another, with all the previously mentioned categories and cover most of the regions listed. Those maps which have a large series of sheets associated with them are designated below by their call numbers, as found in the Map and Geography Library. The number of associated sheets is also given.

	Call Number	Number of sheets
Southeast Asia	G8000s 250 .U5	124

. . . . continued

	Call Number	Number of sheets
Philippines	G8060s 250 .U5	60
	G8060s .C5 1963 .P4	8
	G8062 50 .U51	39
	G8062s .C45 25 .U5	116
	G8062s .L8 50 .U5	218
	G8062 .M5 25 .U5	175
French Indo-China	G8010s 50 .I5	83
	G8000 100 .I5	236
	G8010s 100 .G7	5

. . . . continued

	Call Number	Number of sheets
	G8010s 400 .I5	24
Malaya	G8030s 63 .G7	68
	G8013 .V5 500 .U5	12
Burma	G7720s 250 .U5	45
	G7720 253 .G7	63

The second most important source of maps was in the Map Library of the Department of Geography located in Davenport Hall. This library (maintained and managed by a graduate student) yielded 68 maps. These maps are both regional and topical in nature. They are all single sheet maps; no series of sheets are associated with any of them. Unlike the fine quality and physical condition of the maps in the Map and Geography Library collection, a number of these maps are quite old, show crude cartographical techniques, and, in some instances, are rather worn. All of these maps are wall maps, i.e., they have a wooden frame at the top and bottom, and the majority are large (the smallest ones being 30" x 40"). Practically every topical map covered the whole world; 26 maps dealt with specific regions. Southeast Asia was well represented in this collection.

The number of appropriate maps possessed by the faculty of the Department of Geography at the University was minimal. Professor R. W. Armstrong was the only faculty member with applicable maps, and the number of different sources was less than ten. The maps dealt with climate and vegetation on a world-wide scale; there were no regional maps available.

Search of a number of other departmental libraries did not produce a large quantity of suitable maps for this project. A list of the searched libraries includes: Agriculture, Biology, Chemistry, Geology, Veterinary Medicine, State Natural History, State Water Survey, and The State Geologic Survey. A search at the Agronomy Department office in Turner Hall yielded a number of high-quality soil association maps of the Quadri-County area. It was discovered that a number of these departments have large quantities of textual material and literature available that would be helpful in the MOD project, but, as stated, there were relatively few applicable maps.

5. Conclusion All applicable maps of the size and quality needed for this project have been sought out their information recorded, xeroxed, and sent to the MOD project headquarters in Washington, D.C. All sources of maps have been checked, including different national atlases. The only possible remaining source of cartographical material would be maps contained in monographs, but they are of questionable value for the project. Most of these maps would be not larger than page size, and the amount of detail and degree of coverage would, of course, be limited.

Most of the maps (especially those with a number of associated sheets) in the Map and Geography Library would be well-suited for detailed plotting of data; those maps found in the Geography Library of Davenport Hall would be well-suited for plotting gross data.

continued next page

Search of a number of other departmental libraries did not produce a large quantity of suitable maps for this project. A list of the searched libraries includes: Agriculture, Biology, Chemistry, Geology, Veterinary Medicine, State Natural History, State Water Survey, and The State Geologic Survey. A search at the Agronomy Department office in Turner Hall yielded a number of high-quality soil association maps of the Quadri-County area. It was discovered that a number of these departments have large quantities of textual material and literature available that would be helpful in the MOD project, but, as stated, there were relatively few applicable maps.

5. Conclusion All applicable maps of the size and quality needed for this project have been sought out, their information recorded, xeroxed, and sent to the MOD project headquarters in Washirgton, D.C. All sources of maps have been checked, including different national atlases. The only possible remaining source of cartographical material would be maps contained in monographs, but they are of questionable value for the project. Most of these maps would be not larger than page size, and the amount of detail and degree of coverage would, of course, be limited.

Most of the maps (especially those with a number of associated sheets) in the Map and Geography Library would be well-suited for detailed plotting of data; those maps found in the Geography Library of Davenport Hall would be well-suited for plotting gross data.

continued next page

ENVIRONMENTAL-FACTOR MAP INFORMATION FORM

(mark all applicable boxes)

1. Scope of map:
   - World (entire)
   - World (major section; continent, ocean) specify:_____)
   - S.E. Asia (entire)
   - Thailand
   - Malaya
   - Other (specify:_____)
   - Mid-western Illinois (entire)
   - Southern Illinois (including Quadri-Co. area).

2. EXACT title of map:_____
   (i.e., name of environmental factor mapped)

3. Break-down of values of environmental factor mapped - XEROX LEGEND OF MAP AND STAPLE TO THIS SHEET.

4. Method of representing data on map:
   - dot-type symbols
   - alphabetic/numeric symbols
   - contour-type lines
   - shading/patterns (black-white)
   - shading/patterns (colors)
   - other (specify):_____

5. Projection used:
   - equirectangular (plane-chart)
   - Miller cylindrical
   - Mercator
   - homolosine
   - Other (specify:_____
     also note whether:
     longitude meridians are straight or curved lines
     latitude paralels are straight or curved lines
     projection is interrupted, interrupted and condensed, condensed, or non-interrupted and non-condensed

6. Scale of map: 1/_____

Dimensions of map: (cross out non-applicable units) (width)_____ cm. in. X (height)_____ cm. in.

8. Date of publication of map:_____

9. Date(s) of data mapped: _____ (earliest) - _____ (latest)

10. Data compiled by (if different from item 11):_____

11. Bibliographic reference for map: book--author, date, title of book, Publisher and city, page: journal--author, date, title of article, Name of journal, volume, (number), page.

_____

12. Call number of source containing map: _____

_____

13. Call number of map (if different from item 12): _____

_____

14. Physical location of map (or source containing map):

Univ. of Illinois main library

Univ. of Illinois map library

Faculty member's personal library

(specify whose: _____)

AFIP Ash Library

AFIP Geographic Path. - Geographic Zoon. library

Staff member's personal library

(specify whose: _____)

Other location (specify): _____

_____

15. Additional remarks (use reverse side if necessary):

_____

16. This form was completed on:

(Date) _____ (Name) _____

The major primary sources of maps which were found to be most useful to the MOD project are as follows (arranged in alphabetical order):

Aeronautical Chart and Information Center (U.S.A.F.)

American Geographical Society

Coast and Geodetic Service (Dept. of Commerce)

Earth Science Division of the U.S. Army Natick Laboratories (various atlases dealing with climate, insect vectors, etc.)

F.A.O. (especially in relation to crop ecology)

Geological Survey (Dept. of Interior)

National Geographic Society

U. S. Army Map Service, Corps of Engineers

U.S. Navy Hydrographic

U.S. Naval Oceanographic Office

Shistosomiasis — general considerations

Schistosomiasis represents a group of diseases caused by any one of three blood flukes (S. mansoni -- the one with which we have been primarily concerned in our mapping program -- S. haematobium, and S. japonicum). Each of these specific organisms has its own (geographic) distribution pattern, although they overlap.

As is shown in Figure A-2 (taken from Pathology of Tropical Diseases, by Ash and Spitz), these organisms have a rather complex life cycle, and each requires an intermediate host -- a particular species of snail -- in order to fulfill a vital stage of their development. The infected human being releases (excretes) eggs either in his feces or urine. If these eggs are discharged into fresh water, they hatch as free-swimming organisms (miricidia). Then, if they are able to reach the appropriate species of snail within a short time, they infect the snail (producing disease) and develop into a larval stage. At this stage they break out of the snail to become a free-swimming form once again (cercariae). If man is exposed to infested water for only a moment or two, the organism can actively force entry through his skin and into cutaneous blood capillaries or venules. Once in the blood stream they are carried to their particular (organ) site of preference for further development. S. mansoni and S. japonicum prefer the veins which supply the colon -- and the colon and rectum are primarily involved; S. haematobium concentrates in the blood vessels of the urinary bladder -- and the bladder is the primary site of disease. Once the larval organisms reach their maturity, they begin to lay eggs, and it is these eggs which incite a marked chronic inflammatory reaction. Major areas of damage (depending upon the type of organism) are the colon, the liver, the urinary bladder (and, secondarily, the kidneys).

In most instances (most patients do not receive adequate treatment), the disease is a very very chronic one, lasting for many years. It is apt

to become continually worse because continual re-exposure and re-infection ("super-infection") adds continually to the number of infecting parasites.

Schistosomiasis is a very important cause of death, especially among infants. It is an even greater cause of morbidity, and it depletes the energy of millions of people in the world, with very profound socio-economic consequences. Because the life cycle of these organisms are dependent upon an intetmediate host which, in turn, requires a suitable kind of fresh water, etc., etc.) this is one of the diseases which can easily be introduced into a new area -- or eliminated from an old area -- by changing the ecology.

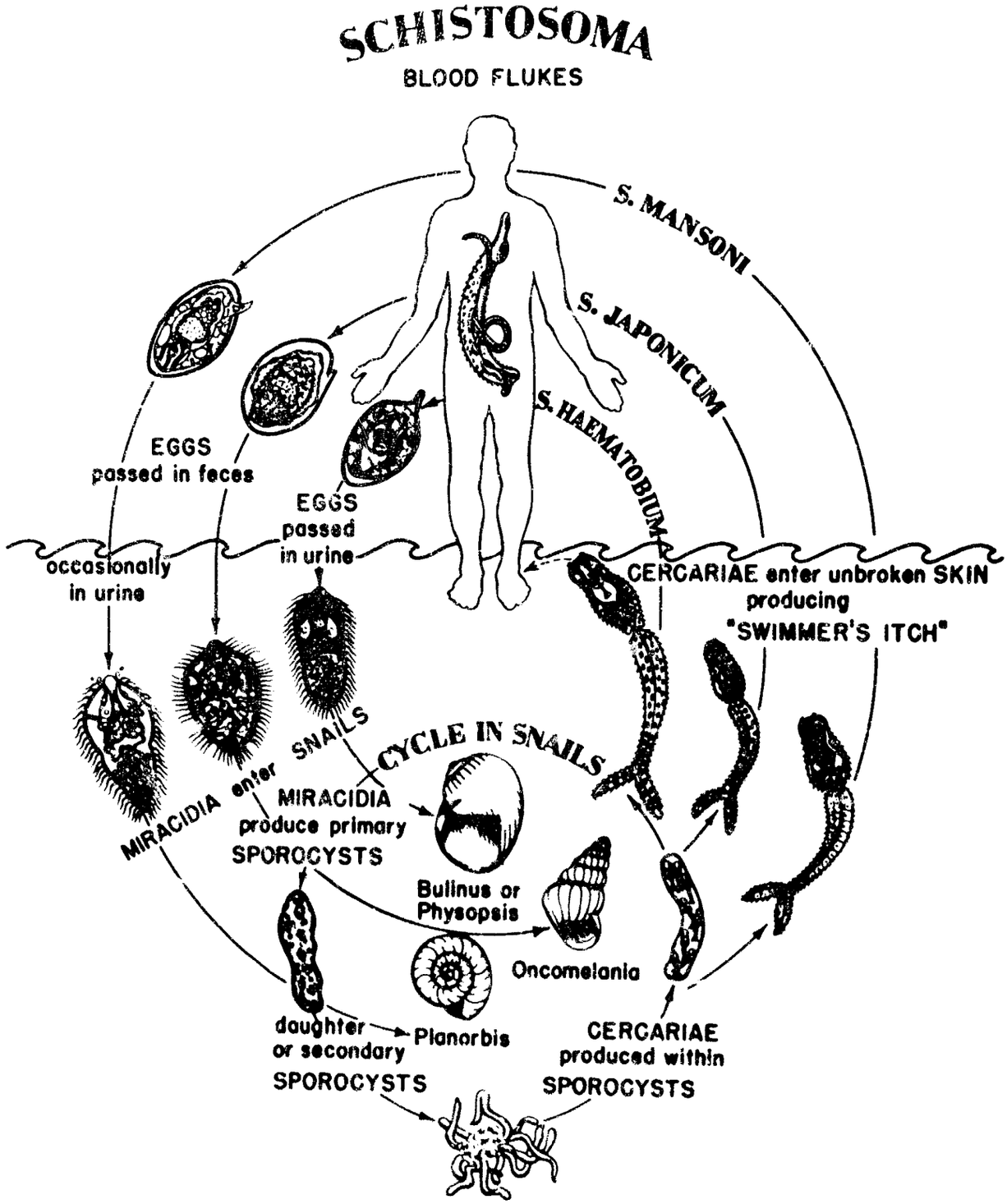


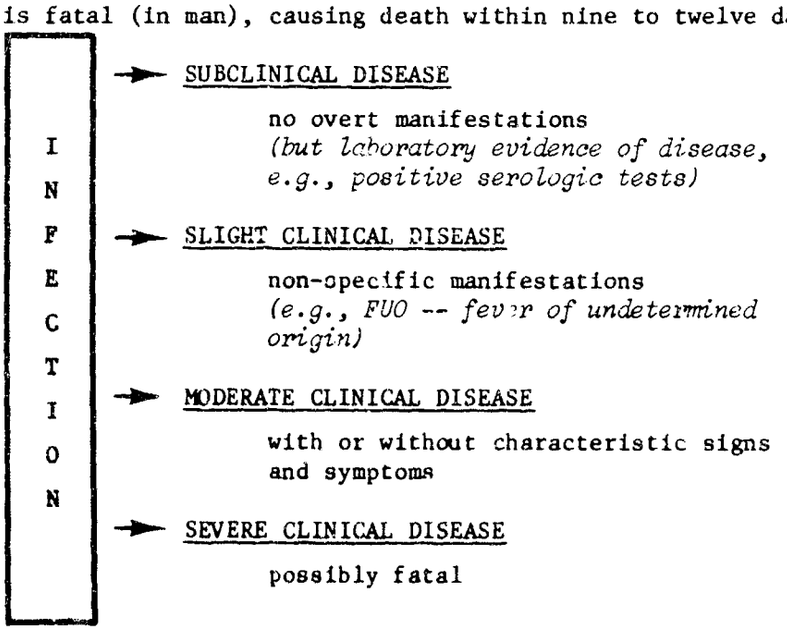
Figure A-1

Phyllis Smith, 1944.

From Pathology of Tropical Diseases by Ash, J.E. and Spitz, S. Saunders Co., Philadelphia, 1945; AFIP neg. #55-1841.

Leptospirosis — general considerations

Leptospirosis (known by many other names, including *Weil's disease*) is an important acute infectious disease of man and animals (i.e., a zoonoses) that is worldwide in its distribution. The disease was first recognized in human beings in 1886. It varys greatly in its manifestations (as shown in the adjacent figure). The severe disease, in man, affects many organ systems, particularly the liver and the kidneys. Occasionally the disease is fatal (in man), causing death within nine to twelve days as a rule.



The causative organism belongs to the genus Leptospira of the family Treponemataceae. Although there is only one species pathogenic for man (L. interrogans), there are many many "serotypes", virtually identical in form, but differentiated on the basis of their antigenic characteristics. Since each of the various serotypes has its own peculiar ecologic characteristics, it is important (epidemiologically) to determine the specific serotype responsible for a given infection.

The leptospiral organism is usually transmitted to man by food or drink that has been contaminated by the urine and/or excreta of rats or other

rodents, or from immersion in water which has been contaminated by some animal reservoir. Workers in sewers, irrigation ditches, rice fields, docks, and abattoirs, are at high risk.

Clinically, (in the severe case), the onset is sudden with high fever, headache, and generalized body pains. Depending upon the severity, renal and hepatic involvement may lead to uremia and jaundice. Ordinarily, the disease runs a course of three to four weeks in human beings, with a mortality rate of about 5% (of the clinically recognized cases).

Leptospirosis was chosen as the principal disease to study by the MOD project team because there are many known animal reservoirs (see Figure A-1), the amount and nature of surface water is an important ecologic factor in maintaining sources of infection, and persons of certain vocations are at high risk, etc. etc. Furthermore, there is still a great deal to be learned about leptospirosis. Two of the (13) recommendations for research in a recent (1967) Report of a WHO Expert Group were as follows:

- *Further studies of the ecology of reservoir hosts are necessary, particularly in regions where environmental conditions are rendered unstable by human activity.*

- *Results of ecological studies should be applied to the forecasting of cycles of infection and to systematic surveillance programmes on which prevention of outbreaks in man could be based.*

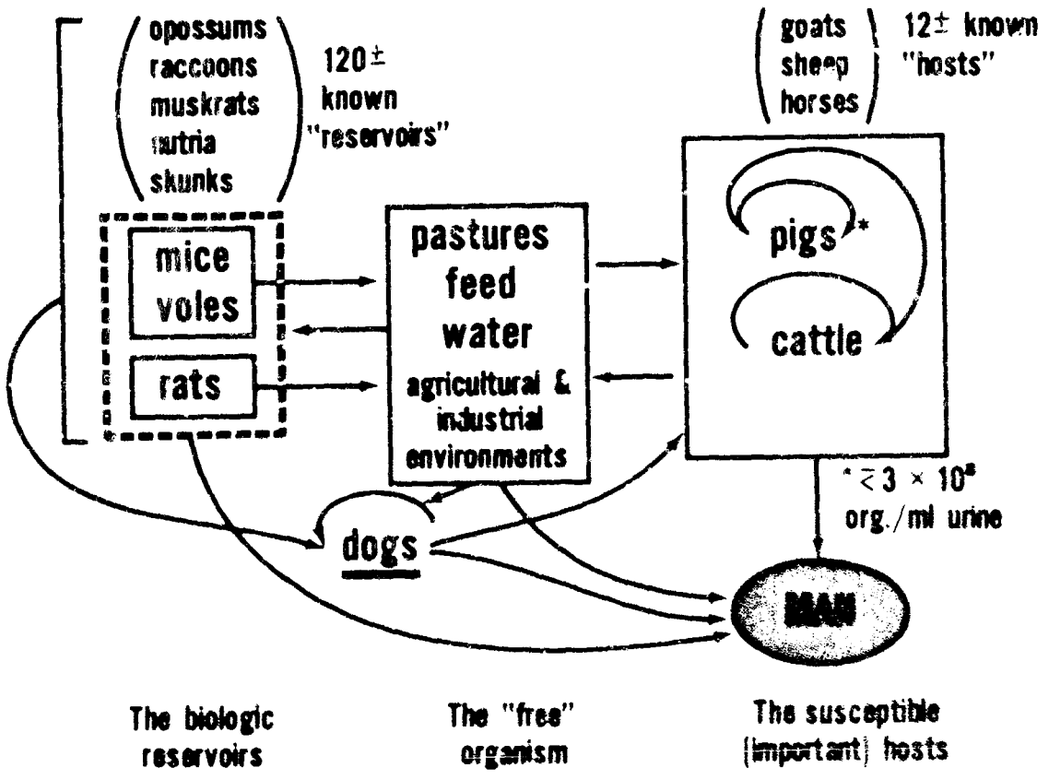


Figure A-E A schema showing relationships among man and animals -- and their environment -- as they pertain to the ecology of leptospirosis.

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) The Universities Associated for Research and Education in Pathology (UAREP)	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

THE GEOGRAPHIC DISTRIBUTION OF INFECTIOUS DISEASES
[The Mapping of Disease (MOD) Project]

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Final

5. AUTHOR(S) (Last name, first name, initial)

Pugh, H. C. Morenoff, J.
Gaffey, R. J. Richmond, W. L.
Sidley, J. D. H.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
31 August 68	432	295

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
DA-49-092-ARO-130 b. PROJECT NO.	None
c. d.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) Program Code No. P6G20-25

10. AVAILABILITY/LIMITATION NOTICES

This Document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
Contract and Technical Monitoring Agency: Life Sciences Division; OCRD; Dept. of the Army, Wash., D. C.	Advanced Research Projects Agency; Dept. of Defense; Wash., D.C. 20301

13. ABSTRACT The occurrence and character of disease are greatly affected by many environmental factors. It is very difficult to determine which of these bear causal relationship to disease, and under what conditions, by present methods of inquiry. This project was designed to develop methodology for rapid and effective searches for important interrelationships among the large variety of potential causal factors of any given disease, in a geographic and time context. An immediate objective was to produce maps showing disease distribution. The major (ultimate) objective was to develop a system whereby: (1) input data relating to disease and environment could be characterized for effective storage and retrieval *in context* by a computerized system which, (2) using these data, could relate meaningfully the prevalence, incidence, character, and (geographic) location of disease to a variety of direct and indirect (environmental) causal factors, and (3) output the information directly in map form, as well as in tables and graphs.

This report is a detailed account of the analysis and design of a computer system that can accomplish the above objectives, with [illegible] illustrations of computer produced disease maps that attest to the validity of the basic proposition as well as the efficacy of the proposed computer system. The project had to be terminated before Phase III, implementation, could be achieved, thus the (majority of) programs necessary to operate the system have not been produced. Data sources, extraction, and preprocessing are considered in detail, including the development of a new data structuring system pertinent to disease ecology, and a list of "disease factors". Individual sections of the report deal with output analysis and output usage.

DD FORM 1 JAN 64 1473

Security Classification

14. KEY WORDS		LINK A		LINK B		LINK C	
		ROLE	WT	ROLE	WT	ROLE	WT
pathology	zoonoses						
geographic pathology	schistosomiasis						
medical geography	leptospirosis						
disease maps	etiology						
medical ecology	information storage						
epidemiology	and retrieval						
ecology	data structuring						
geography	terminology						
mapping	data extraction						
cartography	data collectic.						
disease	data sources						
infectious disease	disease factors						
infection	reservoirs of infection						
	vectors of disease						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS), (S), (C), or (U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.