Climatic Data Banking and Analysis Using an Information Retrieval System

R. G. BARRY,[1] R. C. BRILL[2] AND G. F. ESTABROOK[3]

(Manuscript received 5 February 1973, in revised form 4 September 1973)

ABSTRACT

The TAXIR information retrieval system, originally developed for taxonomic research, is described in the context of its application to climatological data. Data banks for four mountain stations in Colorado have been established and analyzed using TAXIR and a package of statistical routines. The procedures and their cost effectiveness are evaluated.

1. Introduction

Analysis of climatological data may entail routine reduction of observed records for standard periods (monthly, annual) or, alternatively, computations of specified climatic conditions. Analysis of the former type are computationally trivial whereas those of the latter usually require special purpose programs or, at least, the arrangement of the data in a format suitable for existing "package" programs.

This paper describes an information management system offering great flexibility in analytical procedures, and illustrates its application to climatological data banks.

2. The TAXIR System

The TAXIR (*Tax*onomic *I*nformation *R*etrieval) system was originally developed for use in systematic biology. The researcher has at his disposal the language of Boolean algebra for specifying the subset of his data that he would like to retrieve. TAXIR translates the researcher's Boolean expression into an internal form and proceeds at once to perform the necessary Boolean arithmetic directly on the stored data. This calculation isolates the subset of interest from the entire body of stored information. Because TAXIR retrieves by calculation rather than by traditional searching techniques, the retrieval is extremely rapid. Moreover, the data are stored in a highly compact form, which not only makes them receptive to Boolean calculation, but also economizes the use of machine storage space (Estabrook and Brill, 1969).

The TAXIR system involves the following procedures:

1. Building a data bank.
2. Correcting and revising a data bank.
3. Retrieving desired portions of a data bank.

Each element of a data bank is referred to as an *item*. For our purposes, items are commonly station-days of climatic data. The information comprising an item consists of *descriptors* such as temperature, humidity, wind speed, day, month, year, etc. Each descriptor is divided into *descriptor-states*. In quantitative descriptors these states will consist of a range (e.g., in the descriptor WIND SPEED the states may constitute the number set from 0 to 100 in steps of 1, measured in MPH). In qualitative descriptors, which played no part in the climatological application, the states will merely be a list of terms culled from the data. Items, descriptors and descriptor-states are defined by the researcher with terms of his own choosing.

The range of a descriptor may in some cases have to be guessed or chosen with sufficiently wide limits to allow for future extremes. However, the range of each descriptor should be specified as conservatively as possible, since the amount of computer storage space required for a particular data bank depends primarily on the number of items, descriptors and descriptor-states which are defined.

It may be necessary to correct or revise information in a data bank once it is established. The TAXIR language includes a "correction" statement for removing errors in individual items or sets of items; a "deletion" statement for deleting spurious descriptor-states and another for deleting unwanted items; and a statement for defining additional descriptors. Additional items can always be incorporated if their descriptor-states fall within the ranges earlier specified and space is available in the computer memory.

Data retrieval is achieved by posing a query which incorporates a Boolean expression, but which nevertheless resembles plain English and is based on the researcher's own terminology. A query specifies logically

[1] Institute of Arctic and Alpine Research, University of Colorado, Boulder 80302.

[2] Coolie Laboratory, University of Michigan, North Campus, Ann Arbor 48104

[3] Department of Biology, University of Michigan, Ann Arbor 48103.

TABLE 1. List of descriptors in daily banks.

Descriptor	Descriptor states	Description
STATION	6	
YEAR	(15)*	
MONTH	12	
DAY	31	
TMAX	127	Daily maximum temperature (°F)
TMIN	126	Daily minimum temperature (°F)
*TEMP V	3	Temperature data reliability (valid, estimated, doubtful)†
TMEAN	121	Daily mean temperature
TRANGE	63	Daily (maximum−minimum) temperature range
RHMAX	101	Daily maximum relative humidity (%)
RHMIN	101	Daily minimum relative humidity
RH V	3	Humidity data reliability
RHMEAN	101	Daily mean relative humidity
RH RANGE	101	Daily (maximum−minimum) relative humidity range
*RAD	101	Daily solar radiation total (10 ly)
*RAD V	3	Radiation data reliability
*PRECIP	511	Daily precipitation (0.01 inch)
*PRECIP V	3	Precipitation data reliability
*WIND	127	Mean daily speed (mph)
*GUST	201	Peak gust (mph)
DEN T	62	700-mb Denver temperature, 1200 GMT (°C)
DEN RH	101	700-mb Denver relative humidity, 1200 GMT (%)
DEN DIR	37	700-mb Denver wind direction (10°)
DEN VEL	63	700-mb Denver wind speed (m sec^{-1})
FL	351	Height of Denver freezing level (10 m)
UA 1	31	700-mb statistical type (heights)
UA 2	31	700-mb statistical type (height anomaly)

* Part 2 only.
† Daily part 1 has TMAX V and TMIN V instead of TEMP V.

some properties of items. TAXIR locates the subset of items with these properties and prints the number of items in the subset, the number of items in the data bank, the proportion of the total number of items represented by the subset, a conveniently formatted listing of the states of specified descriptors displayed by the items in the subset, and, if requested, a magnetic tape or disc file of these data.

Although TAXIR was not originally conceived as a tool for statistical data processing, it has now been suitably modified for this purpose since the tape file provides input to a package of statistical programs.

A user's primer for the TAXIR system has been prepared by Brill (1971). The system is operative on the CDC 6400 at the University of Colorado and on the IBM 360/67 at the University of Michigan and at the U. S. Department of Agriculture, Plant Introduction Station, Washington State University, Pullman.

3. Climatic data banks: An illustration for the Front Range, Colorado

Climatic records have been collected at four elevations between 2195 and 3750 m on the east slope of the Front Range in Colorado since October 1952. Four stations have been maintained on ridge top locations from just west of Boulder almost to the Continental Divide (at approximately 40N, 105°22′W to 105°37′W). For the 12 months October 1952–September 1953 records were also taken at north-facing, south-facing, and valley sites in each of the four elevation zones. The basic data were tabulated for 1952–64 (Marr, 1967; Marr *et al.*, 1968a, b).

Four TAXIR data banks have been established on tape file as follows:

1. Sixteen-station bank for October 1952–September 1953 (5,840 records).
2. Daily bank, part 1; four stations, 1952–64 (17,520 records).
3. Daily bank, part 2; four–six stations, 1963–70 (11,680 records).
4. "Weekly" bank; four stations, 1952–64 (~2,400 records).

The item in the first three banks is a station-day. In the weekly bank it is a station-"week". The irregular servicing interval necessitated that the duration since the last servicing be included as a descriptor. The daily bank was divided into two parts due to space limitations in the computer storage and because data for several parameters were available only in the latter part of the record. These two units can be examined jointly, however, by concatenating TAXIR-generated tape files with a special purpose program.

Table 1 illustrates the descriptors included in the daily banks covering the period 1952–70 (with space for additions through 1976). In addition to the observed elements, the bank includes parameters such as mean and range derived in a conversion program during the formatting of data prior to its entry in the bank. Since temperature, humidity, precipitation and radiation data were abstracted from chart records, descriptors relating to their reliability are included. There are also descriptors of the synoptic conditions relating to 700-mb rawinsonde data at Denver and to classifications of the daily 700-mb height field and its deviation from 1952–70 monthly mean pattern. The classifications were obtained by using the correlation techniques suggested by Lund (1963) based on heights at 31 grid points covering the sector 20–60N, 95–125W. This analysis will be discussed elsewhere; the details are given in Barry (1972).

4. Application in retrieval and analysis

The retrieval of specified subsets of data is carried out by writing TAXIR statements containing Boolean expressions. The three Boolean operators "complement" (NOT), "intersection" (AND), and "union" (OR) logically relate properties of the data to determine the desired set. Basically, a property comprises a descriptor name together with one of its states or a range of states.

A typical query statement might be:

PRINT: STATION, MONTH, (DAY, TMAX, TMIN, TMEAN, TRANGE) FOR ALL STATION DAYS WITH TMAX, FROM −21 TO 32 AND YEAR, 1965

Boolean expression

The response to the above query would appear as follows:

NO. OF ITEMS IN QUERY RESPONSE=655
NO. OF ITEMS IN THE DATA BANK=2250
PERCENTAGE OF RESPONSE/TOTAL DATA BANK=29.11

STATION 1				
JAN				
1	28 DEG F	13 DEG F	20 DEG F	15 DEG F
25	28 DEG F	15 DEG F	21 DEG F	13 DEG F
26	28 DEG F	14 DEG F	21 DEG F	14 DEG F
28	32 DEG F	23 DEG F	27 DEG F	9 DEG F
FEB				
1	27 DEG F	9 DEG F	18 DEG F	18 DEG F
2	31 DEG F	8 DEG F	19 DEG F	23 DEG F
10	20 DEG F	5 DEG F	12 DEG F	15 DEG F
11	18 DEG F	−2 DEG F	8 DEG F	20 DEG F
12	20 DEG F	4 DEG F	12 DEG F	16 DEG F
15	21 DEG F	2 DEG F	11 DEG F	19 DEG F
23	20 DEG F	−2 DEG F	9 DEG F	22 DEG F
MAR				etc.

The output tabulates the four temperature descriptors for dates in the year 1965 with a maximum temperature within the specified range sequentially at each station. The descriptors enclosed in parentheses in the query statement are printed on the same line.

The types of problem which are ideally suited to the TAXIR retrieval capabilities are, in order of increasing complexity, as follows:

1) Selection of dates on which extreme values occur or the determination of extremes on specific dates. In each case the response to the query will be small if a reasonable lower cutoff can be estimated.

2) Tabulation of the number of occasions of a specified event (or range of conditions) or a listing of the actual dates and locations of such occurrences. The event may, of course, be a complex combination of descriptor-states; for example, it may be required to know the number of days with maximum temperature within a specified range, no precipitation, and wind speed less than a given value. This could be requested by the following:

HOW MANY DAYS WITH WEATHER STATION, 1 AND TMAX, FROM 70 TO 100 AND PRECIP, 0 AND WIND, FROM 0 TO 15

Another query designed to print the values of solar radiation on dates with the conditions specified above might read:

PRINT: YEAR, (MONTH, DAY, RAD) FOR ITEMS WITH WEATHER STATION, 1 AND TMAX, FROM 70 TO 100 AND PRECIP, 0 AND WIND, FROM 0 TO 15

This type of query provides the material for complex n-way contingency tables.

3) The printing of climatic tables for an entire data bank. The printout may be ordered on any selected leading descriptor.

Before requesting printout of a large subset, such as an entire data bank, it is advisable to ensure that the proper output format is being obtained by posing a sample query that is certain to retrieve only a few items. This will obviate the risk of erroneous and perhaps costly computer runs.

In addition to the above TAXIR queries, standard statistical analyses can be performed on the data by generating a tape file. This file may be a temporary special purpose one, or it could include the entire bank. If the data bank is open-ended a file of new items can be generated when required and concatenated with the original one.

Time-related subsets of data are not generally accessible through Boolean expressions. However, we may be interested, for example, in weather conditions at the same, or a different station, on days preceding or following a specified event. A small computer program (Time File Generator) has been developed, in connection with a separate study of avalanche occurrences, to cope with this problem. A chronologically ordered file (Tzero) is first generated by TAXIR from the data bank of interest. The Time File Generator program is then used to create another file (Tdiff) in which for every item in Tzero there is a corresponding item in Tdiff with a date differing by some fixed number of days (or other standard unit time interval). More than one Tdiff file can be generated and stacked on a single tape. These files can then be processed by the programs in the statistical package.

5. Cost evaluation

A cost analysis was maintained during the project as a basis for assessing the cost effectiveness of the TAXIR approach. Table 2 summarizes the major cost categories. A brief explanation of some of the items follows:

TAXIR development—includes modification of the then existing programs and addition of some new functions.

Statistics programs—modification and testing of the programs incorporated in the package and development of some new ones.

Pressure data programs—conversion programs for the selection of the required data from tapes of 700-mb heights and for the correlation analysis and grouping.

Data conversion and data banks—includes preparation of back-up tapes for the banks.

Pressure data analysis—the correlation and grouping analysis for 700-mb height and anomaly patterns in each month 1952–70.

TAXIR queries—where the output was in a tape file format and subjected to further analysis, these costs are included under "statistical analyses".

The proportionate cost of "unsuccessful" computer runs is considerably greater in the development stage (41%) than in the production stage (29%). However, incomplete runs made during the debugging phase of development, which have not been differentiated in the accounting, nevertheless provide essential information in developing a satisfactory program and should not be regarded as wastage. It may be noted also that TAXIR queries are virtually foolproof during the production stage. In the time-sharing mode with a remote terminal (available at the University of Michigan), a system of interactive error recovery and correction has been implemented on the TAXIR system.

Comparison of permanent file storage and tape indicates that the latter is more economical unless numerous jobs are to be carried out. The TAXIR program occupies 250 physical record units (PRU's), or approximately 2000 punched cards. Permanent file storage on the CDC 6400 costs $25 per month (at $0.10 per PRU) whereas tape input or output costs only $0.40 per 1000 PRU. Allowing for charges of $0.50 per job to mount a data tape and a proportion of the cost of the tape, at least 30 jobs per month would have to be run to reach the break-even point. For significantly larger programs or data the advantage is decisively for tape storage.

Projections of the increase in storage and operating costs when the amount of data is augmented by an order of magnitude or more are very desirable but, due to the variety of factors which come to bear on this question, such estimates would be quite speculative. The present description of one moderately large implementation of the system affords a basis for estimating costs and efficiencies of comparable sized banks on similar computing equipment. If the size of the bank were increased by two orders of magnitude (to 10 million station-days of data), economies of very large scale on the one hand and limitations of computer hardware on the other, demand redesign of the system in ways that conceivably may be less efficient for medium-scale implementation. TAXIR has been implemented on the IBM 360/67 (duplex under MTS) at the University of Michigan to meet the needs of several medium-to-large-scale users and some design changes dictated by changes in user needs and operating facilities have been made. Detailed cost data resulting from some large-scale implementations at the University of Michigan should be available in the near future, but it appears that in this case the cost per user is reduced substantially by the opportunities for cost sharing.

In order to compare TAXIR with a conventional mode of analysis we may postulate that the latter would comprise storage of the climatic data and a synoptic catalog on card (or magnetic tape) files, and the use of

TABLE 2. Summary of costs.

	Computation costs					
	Productive runs	Unsuccessful runs	Total	Manpower costs	Other costs	
Development						
TAXIR development	$ 313	$ 92		$ 2,600	Data abstraction	$ 300
Statistics programs	197	287		4,300	Punching and verifying	1,600
Pressure data programs	277	177		300	Upper air data	615
					Tapes, cards	250
	$ 787	$ 556	$1,343	$ 7,200	Keypunch rental	1,800
Unspecified development and testing			890			
			$2,233			
Production						
Data conversion, data banks, etc.	$ 698	$ 556		$ 2,000		
Pressure data analysis	1,728	599				
TAXIR queries	349	28		11,413		
Statistical analyses	878	283				
	$3,653	$1,466	$5,119			
			$7,352	$20,613		$4,565

existing computer routines for statistical analysis (given some special purpose programs to manipulate the data). The costs of such an analysis are not easily estimated directly but this question can be approached, using Table 2, by considering what features would be different from those associated with the TAXIR study.

The expenditures in the development stage (33% of the total costs for computation and consultancy) should not be included as they stand, since these represent an investment which is already being realized in other projects. If the systems were to be used, for example, in another four projects it would be reasonable to include 20% ($1,900) of these costs. The costs listed as "Other" in Table 2 can be ignored as they would be incurred in both a TAXIR and a non-TAXIR approach. Manpower costs in the production stage would also be likely to be very similar in either type of approach. The manpower cost ($2,000) involved in the conversion and formatting for the data banks would be balanced by the cost to process the data to obtain daily mean values and ranges and incorporate these on the daily data cards (estimated at $1,000). Examining the computation costs in the production stage, we will assume that 1) the pressure data analysis would be the same (and this would provide the synoptic catalog), 2) contingency analysis would have to be substituted for the TAXIR queries with a higher cost for a standard linear search through the data files, and 3) the statistical analyses would each require a special formatting of the data. It is the last two items which are the most difficult to cost. The principal difference, assuming that a sort/merge program was available or could be readily developed, would occur in the time required to read in the TAXIR data banks as against a binary data tape. Experience shows that for the entire 18-year record the former costs about $0.35, the latter about $5.00 on the CDC 6400. Depending on the amount of processing required, the cost ratio is between 1.3–2.8 times greater for the non-TAXIR analysis. Assuming a factor of 2, the contingency and other statistical analyses would therefore cost ~$3,000. Using these assumptions we can cost the non-TAXIR approach and make a comparison with those of Table 2 with the modifications indicated. This is shown in Table 3. The costs are remarkably close for either method.

TABLE 3. A comparison of actual costs and estimated costs for a non-TAXIR approach.

		TAXIR	Non-TAXIR
Development:	pressure data	$ 454	$ 454
Development:	computation	356*	—
	consultant	1,680*	300
Other costs:		4,565	4,565
Production:	data banks	1,254	1,000†
	pressure data	2,327	2,327
	statistical analyses	1,538	3,000
	manpower	13,413	11,413
		$25,787	$23,059

* Assumes 20% of cost of TAXIR and statistical programs.

† Cost to derive means and ranges; data card punching in Other costs.

There are, of course, additional advantages of the TAXIR system which have not been considered in the above. The daily data bank can be readily extended as new observations are collected and new descriptors can be added. More important, the cost of further TAXIR queries to any interested user will be much less than that of a comparable special purpose analysis. For a permanent data bank which is likely to be of interest to scientists in a variety of disciplines this is a major consideration.

6. Concluding remarks

In the climatological project that has been outlined, permanent data banks and a flexible retrieval and analysis system have been established. The TAXIR system has been intensively evaluated and it is shown that the cost of developing and using this procedure is close to that of a conventional analysis. However, the established system will greatly reduce the cost of subsequent analyses with these data as the banks are extended. Moreover, much of the investment in the system is already being realized in connection with other projects.

Acknowledgments. The original development of TAXIR was supported by the National Science Foundation under Grant GN-656 to D. J. Rogers, University of Colorado. The system was designed and programmed by R. C. Brill and G. F. Estabrook. Modifications to the system and its application to the climatic banks were supported by the Atmospheric Sciences Section of the National Science Foundation (GA-15528) and programmed by R. C. Brill. Much of the routine operation was carried out by Mrs. M. Eccles aided in various ways by Miss W. A. R. Brinkmann and Mrs. Jill Williams. The statistical package was developed by Mr. Z. Little, Institute of Behavioral Science, University of Colorado.

REFERENCES

Barry, R. G. 1972: Climatic environment of the east slope of the Colorado Front Range. *Inst. Arctic Alpine Res., Occas. Papers*, No. 3, University of Colorado, 206 pp.

Brill, R. C. 1971: The TAXIR primer. *Inst. Arctic Alpine Res., Occas. Papers*, No. 1, University of Colorado, 72 pp.

Estabrook, G. F., and R. C. Brill, 1969: The theory of the TAXIR accessioner. *Math. Biosci.*, **5**, 327–340.

Lund, I. I. 1963: Map-pattern classification by statistical methods. *J. Appl. Meteor.*, **2**, 56–65.

Marr, J. W. 1967: Data on mountain environments I. Front Range, Colorado, sixteen sites, 1952–53. *Univ. Colo., Ser. Biol.*, No. 27 (Boulder), 110 pp.

——, A. W. Johnson, W. S. Osburn and O. A. Knorr, 1968a: Data on mountain environments II. Front Range, Colorado, four climax regions, 1953–1958. *Univ. Colo., Ser. Biol.*, No. 28 (Boulder), 170 pp.

——, J. M. Clark, W. S. Osburn and M. W. Paddock, 1968b: Data on mountain environments III. Front Range, Colorado, four climax regions, 1959–1964. *Univ. Colo., Ser. Biol.*, No. 29 (Boulder), 181 pp.