A FISHERIES DATA BASE UTILIZING SAS
Dan Watson, George Roth, and Leanna Pristas
Texas Instruments Incorporated

ABSTRACT

In order to completely describe the biological activity in a body of water, it is often necessary to perform a wide range of sampling efforts. The diverse nature of the data resulting from activities such as trawl sampling, benthic sampling and mark/recapture studies presents serious problems in designing a data base which is both flexible and easily used. The Ecological Services Department of Texas Instruments Incorporated has designed and implemented a data system for fisheries data, based upon SAS data sets and SAS software. By making use of the ability of SAS to emulate a hierarchical file structure, it is possible to split the data into "levels" of data resolution, which can easily be recombined in a number of different ways to fulfill report and analytical requirements. The system naturally leads to the creation of a "data dictionary", which serves as a common base of communication between biologist and programmer. Significant reductions in cost and time for report generation and training have been realized over previously used FORTRAN based systems.

INTRODUCTION

Since 1972 the Ecological Services Department of Texas Instruments Incorporated (Ecological Services) has engaged in a wide range of aquatic sampling programs. The data collected from these programs have been utilized for measuring base line biological activity and monitoring the impact of changes in aquatic environments. At the onset sampling plans and data collection methods were designed to fit the needs of a client's study and the nature of the environment in which the study was to be performed. As Ecological Services developed, it became clear that a data processing system based on FORTRAN software accessing card-image data files was not flexible enough to handle the wide range of data types and analyses encountered. The processing system's complexity and inflexibility began to impact studies from the Operations level (field and laboratory personnel) through the Data Center (programmers and data clerks) to the Scientific Staff (biologists and statisticians). This impact was most noticeable on large, long-term studies which required that the system adapt to changes as a function of methodological improvement or sampling design.

Before 1979, SAS had been used by Ecological Services for statistical analyses, with only an occasional data processing task done using SAS software. The experience gained from these occasional efforts convinced us that a number of processing problems could be alleviated by conversion of the primary data processing system to one based on SAS software and SAS data sets. It was expected that this conversion would result in improvements in documentation, training, and system performance (Wilson 1981).

The nature of fisheries data implies a hierarchical file structure. Caviar (1978), Harrell (1978), and others have shown that the SORT/MERGE capabilities of SAS allow it to simulate a hierarchical structure using multiple rectangular files. Using SAS in this manner also allows for improvement in file maintenance and flexibility of analysis over FORTRAN based systems with hierarchical card-image data files.

This paper describes the file structure which resulted from the incorporation of the above concepts, along with the accompanying data entry, QA/QC and report generation procedures. The flow of the data processing effort can be represented as follows:

\[ \text{Field/Lab} \]
\[ \text{Data} \]
\[ \text{Phase 0 Processing} \rightarrow \text{Rectification Data Entry (ECOKEY)} \]
\[ \text{Phase I Processing} \]
\[ \text{Univariate Audits} \rightarrow \text{QA/QC Reports} \]
\[ \text{Multivariate Audits} \]
\[ \text{Phase II Processing} \rightarrow \text{Standard Reports} \]
\[ \text{SAS Master File} \]
\[ \text{Special Reports} \]

DATA ENTRY

Typical fisheries data coding systems either restrict change, are inefficient and confusing to field personnel or generate escalating processing and handling costs with each year of use. Our approach to this problem involves the reduction of common variables to standard information blocks. Each project selects and utilizes only those information blocks required to support its data input needs. The forms required for input of laboratory data are produced on standard paper or, in
cases where delayed processing is possible, generated by computer. Computer generation of laboratory data forms permits better inventory control, subsampling, and processing.

Actual data entry is performed utilizing software maintained on a TI-990 computer (ECOKET), which also performs preliminary Phase I Audit (data QA/QC). Data entry is designed around the standard information block coding which systematically and functionally breaks processing down into smaller, easier to manage groups. After Phase 0 and Phase 1 processing is completed, the data is passed to an IBM computer for final QA/QC and master SAS file preparation (Phase II). The major features of this system are:

- Optimization of data entry efforts
- User defined limits during QA/QC permits analytical control on a product or effort basis
- A summary of suspect data is generated in a User (non-programmer) understandable form, facilitating error resolution
- QA/QC is completed prior to any report generation acceptance QC is possible

"ECOKEY" (Hill, 1981) is a key-to-disk data entry system designed to run either on a TI-990 minicomputer or an IBM mainframe. It performs limited variable verification and checking upon entry of the data. Data entry formats can be set by the data entry clerk to reduce the number of encoding error and increase the speed of data entry.

DATA AUDITS AND QA/QC

After the data has been keypunched and verified the audits are performed to identify unusual or outlier values so that they can be corrected or validated before the data is mapped into the master data base. The Phase I data audit first compares individual variables with qualitative and quantitative limits set by the user. The limits can be modified easily to account for seasonality or changes in format. The audit generates a report which summarizes suspect data values for resolution and if the occurrence of a particular species is important (an endangered or rare species, for example), it can be summarized and included in the suspect report upon completion of this effort. Data is then passed to the IBM computer for further QA/QC and final processing. The processing consists of multivariable audits where limiting dependencies between variables are compared, based upon test algorithms set up by the user.

After completion of audits, the data is mapped into a SAS data base and a number of "standard reports" can be requested. The report software can be utilized by any project with the required input variables. Some of the reports which are available are summaries of species densities, catch-per-unit-effort, standing crop and fecundity. The development of standard reports permits technical committee review of processing methods to assure "scientific integrity" and that state-of-the-art methodologies are employed. This portion of the system is designed for use by a biologist with no support from a programmer.

FILE STRUCTURE

The data base for each project consists of a group of SAS data sets in which the fisheries data is divided into 8 "levels", based upon processing and use characteristics. Each level of data is stored in a separate SAS data set, with each observation unique and identified by a set of "key" identifiers variables. These variables allow the matching of appropriate observations in different levels. The level designations and their key identifiers are:

<table>
<thead>
<tr>
<th>Level</th>
<th>Key Identifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 0</td>
<td>Project Code</td>
</tr>
<tr>
<td>Level 1</td>
<td>, Study Year</td>
</tr>
<tr>
<td>Level 2</td>
<td>Task Code</td>
</tr>
<tr>
<td>Level 3</td>
<td>Sample No.</td>
</tr>
<tr>
<td>Level 4</td>
<td>Gear Code</td>
</tr>
<tr>
<td>Level 5</td>
<td>Taxa Code</td>
</tr>
<tr>
<td>Level 6</td>
<td>Individual ID#</td>
</tr>
<tr>
<td>Level 7</td>
<td>Subindividual ID#</td>
</tr>
</tbody>
</table>

For any level the set of key identifier variables consists of the same variables used as identifiers in the next highest level, plus a specific level identifier variable. Except for key identifiers no other variables are allowed to exist at more than one level. In addition, no two observations at the same level can have the same values for their key identifiers. These two restrictions greatly simplify the task of file maintenance. The first restriction minimizes the amount of data redundancy and the amount of edits required to correct a file error while the second restriction ensures that there is no ambiguity as to which observation needs to be edited, whether this is to be done using PROC EDIT or UPDATE. It also results in each observation being matched with at most one observation in each of the higher levels, enabling unique level by level margins.

Not all the levels result in actual SAS data sets as levels 0, 1, and 2 are only organizational in nature and contain no actual data. Level 3 is the highest level to have a corresponding SAS data set, with each observation in the data set representing a single field sample. Variables pertaining to the sampling effort itself, such as date, time, location, water condition (as opposed to water chemistry), weather and sample parameters (i.e., boat speed and direction) are stored at level 3 and are not duplicated elsewhere in the data base. The level 4 data set contains information related to the specific sampling device, or gear, used in the sampling effort. Since some sampling efforts require the use of multiple gears in a single sample, there may be more than one observation at level 4 for each observation at level 3. Variables such as gear depth, sample
duration, area swept and sample volume make up level 4 data sets.

The actual catch data is contained in the level 5 data set, with each taxa recovered from the sampling device requiring one observation. Each observation is identified by the taxa's code number, along with the level 4 identification associated with the sampling device. The number of level 5 observations corresponding to a level 4 observation is random, with a no-catch situation resulting in no level 5 observations at all for that gear. Similarly, a zero catch for a particular taxa is indicated by the absence of an observation at level 5 with that taxa's code value. Variables such as numbers caught by life stage or length class make up a level 5 data set.

Some aquatic studies require information about the physiology of the organisms in addition to the abundance information provided by the level 5 catch data. This information is acquired by taking a subsample of the individuals captured by the sampling device. This data is stored in a level 6 data set, with one observation per individual and variables such as length, weight, age, sex and fecundity. Any additional analysis which results in multiple records for an individual, such as a listing of stomach contents, is recorded at level 7.

In addition to the advantages in data base maintenance already discussed, the division of the data into levels of "resolution" places the organization of the data into a conceptual framework which is easily understood by the scientists who must communicate analysis requests to the programming staff. The scientist can define the data requirements of an analysis in terms of what "levels" of data are required, along with the restrictions on the values of the key identifier variables which determine the subset of the data to be analyzed. The interrelationship between different types of data is also clarified for the programmer, reducing the confusion which often occurs when diverse types of data must be brought together for an analysis. For the most common reports and analyses MACROS have been established which control the way that the various SAS data sets should be merged and establish guidelines for the processing to be used in non-standard analyses.

There are some obvious trade-offs which result from imposing a level structure on the data base, the most serious being that the data in a single level is of little use by itself for analysis purposes. It is necessary to merge multiple data sets in order to generate the simplest of reports and it is sometimes necessary to merge two levels in order to obtain the data contained in a single variable. This is most noticeable in analyses involving sampling efforts which use only a single sampling device per sample. In these cases there is a one-to-one relationship between levels 3 and 4 and the existence of a separate level 4 data set is unnecessary. Still, we feel that the added flexibility of a level structure more than compensates for the occasional awkwardness.

USER'S GUIDE

In order to ensure consistency and efficiency in the use and maintenance of the fisheries data base, a "User's Guide" has been compiled which documents not only the data file structure and its use but also describes the support elements which exist to supplement the data base. A key element in this User's Guide is a "Data Dictionary" which lists all the variables in the data base. Each variable is assigned a unique SAS name and the Data Dictionary lists these variable names along with the variables level assignment, label, definition, units of measurement and list of possible code values. The Data Dictionary also contains a listing of special missing values used in the data base to document the reason for the missing data. This information serves as a focal point between scientists and programmers to ensure that analysis requests are correctly translated into programs. Along with the key identifiers, the information contained in the Data Dictionary uniquely identifies every data element in the data base.

Since there are some types of data which do not result directly from field sampling but are still needed for analysis purposes, a number of SAS "reference" data sets have been created and are documented in the User's Guide. Observational data such as river flow rates and the occurrence of tidal stages, along with taxonomic data and calculated data, such as strata volumes, are stored as SAS data sets and merged with field data when appropriate. A SAS format library is also maintained to allow the conversion of coded data into a more descriptive form at report generation time.

The SAS User's Guide (SAS Inst., 1979) is included as a part of the fisheries data base User's Guide. In addition to describing the SAS processing language, it serves as documentation for the functions and statistical procedures provided by SAS. The data base User's Guide also contains similar documentation to all in-house generated SAS functions and procedures which are available to any user of the data base. The ability to generate functions and procedures in-house allows the continued use of software generated before the conversion to a SAS based system, as well as standardizing and controlling the methods by which complex analyses are generated.

The data sets which make up the data base are documented in the User's Guide by the outputs from PROC CONTENTS which are indexed by a listing of data set names, locations, and descriptions. This information is supplemented by the on-line documentation provided by PROC CONTENTS history listings.

SUMMARY

It has been pointed out (Strand, 1979) that
environmental data requires a data processing system that is flexible without being excessively complex and cumbersome. We believe that the level-structured data base utilizing SAS described in this paper meets those requirements, at least as far as fisheries data is concerned.

In the two years since implementation of this system was begun, Ecological Services has realized significant savings, both in the cost of report generation and the amount of elapsed time involved in the data processing tasks. There has been a 40% reduction in the cost of labor and computer costs and a 50% reduction in labor required for job completion. There have also been savings in the areas of programmer training with a training time 70% less than required under the previously used FORTRAN based system.

At the present time, various methods of interfacing the fisheries data base with water chemistry data are under investigation. We believe that a proper choice of level definitions will result in a water chemistry data base that will meet requirements of both stand-alone water chemistry analyses as well as studies of the relationship between water chemistry and biological activity. Much work is required in this area as well as expanding the fisheries data base to the point where it can handle all forms of aquatic and marine data.

REFERENCES


