# NOAA Technical Memorandum NMFS-SEFC-10

STANDARDIZED DATA CONDENSATION:

A SYSTEMATIC APPROACH TO EFFICIENT FISHERIES AND ENVIRONMENTAL INFORMATION STORAGE AND RETRIEVAL

Larry L. Massey and John E. Hollingsworth

February 1980

U.S. DEPARTMENT OF COMMERCE
National Oceanic and Atmospheric Administration
National Marine Fisheries Service
Southeast Fisheries Center
Miami, Florida 33149

# NOAA Technical Memorandum NMFS-SEFC-10

STANDARDIZED DATA CONDENSATION:

A SYSTEMATIC APPROACH TO EFFICIENT FISHERIES AND ENVIRONMENTAL INFORMATION STORAGE AND RETRIEVAL

Larry L. Massey and John E. Hollingsworth

February 1980

U.S. DEPARTMENT OF COMMERCE
Philip M. Klutznick, Secretary
National Oceanic and Atmospheric Administration
Richard A. Frank, Administrator
National Marine Fisheries Service
Terry L. Leitzell, Assistant Administrator for Fisheries

*This TM series is used for documentation and timely communication of preliminary results, interim reports, or similar special purpose information. Although the memos are not subject to complete formal review, editorial control, or detailed editing, they are expected to reflect sound professional work.*

ABSTRACT

Standardized data condensation is a new concept, providing a flexible, rapid, and relatively inexpensive means of storing, retrieving, and processing fishery and environmental data and facilitating the combination and joint utilization of most fishery data however diverse. Standardized data condensation has been used successfully in the Atlantic menhaden fishery to create a new data base. The data are simpler and less expensive to retrieve and process than the original raw data.

# STANDARDIZED DATA CONDENSATION: A SYSTEMATIC APPROACH TO EFFICIENT FISHERIES AND ENVIRONMENTAL INFORMATION STORAGE AND RETRIEVAL

## INTRODUCTION

In the last few decades the dependence of the scientific field on computers has increased from almost zero to near 100%, mainly because electronic equipment has the capacity to store large amounts of data and computers have the ability to process these data. In fisheries, extensive sampling programs have provided large quantities of data, most of which are stored on auxiliary computer equipment.

In this paper we describe an efficient and relatively inexpensive system for storing and retrieving fishery data. The system, called Standardized Data Condensation (SDC), condenses large data sets into small auxiliary sets that are more readily available and less expensive to process.

Manipulation of SDC data depends on available software and hardware packages. SDC methodology can be used with all digital computer data-storage devices, is extremely flexible with regard to master file formats, and is particularly applicable if month, day, and year of collection are recorded on the master file. SDC can be used for standardizing and condensing input data for data base management systems that specialize in rapid access and retrieval of data (Cohen 1976).

# METHODOLOGY OF SDC

SDC is a two-part logical process. First, annual calendar dates in month-day units are converted into primary fishing units (PFU) (Table 1). Second, pertinent variables within each PFU are summarized. A PFU is one week, the first being January 1-7 of each year. Sets of four PFU's constitute a secondary fishing unit (SFU), which is equal to a lunar month of 28 days. Sets of 13 PFU's constitute a tertiary fishing unit (TFU), equal to a quarter year. Since 365 days cannot be divided evenly by seven, PFU-9 contains 8 days in regular years and 9 days in leap years. Regular calendar months begin on different days of the week and split weeks differently each year. Although PFU, SFU, and TFU units begin on different days of the week each year, an SFU or TFU does not split a PFU week.

Multiple records are condensed to a single entry, one record, in each PFU. For instance, if catch were recorded daily for each vessel, only total catch of all vessels need be recorded for each PFU, thereby reducing the number of records to one. The total SDC file for 30 weeks of fishing by 10 vessels, therefore, would contain only 30 records, rather than a possible 1,500. Each record might contain: PFU number, location of the plant, total catch of all vessels, and the number of days fished by all boats.

The SDC system makes statistics (defined by Schneidewind (1972) as summaries of raw data) available instead of raw data. From SDC statistics one can draw inferences about characteristics of a population or the environment during a PFU week. Estimates of error associated with sampling, such as standard error of the mean, can be computed and stored with PFU statistics if desired.

One advantage of SDC is that statistics can be utilized more readily than raw data. An SFU-month, TFU-quarter, or annual summary can be obtained by summing appropriate PFU summaries, whereas calendar months or quarters cannot be obtained by summing calendar weeks. An investigator can select the time units most appropriate for each analysis. PFU summaries can be used in basic research, in making stock assessments, or in formulating management regulations.

Another advantage is that multiple types of data in condensed files can be easily merged, since the condensed files are all in the same format. Three reference keys are used to merge SDC data sets: year, location, and PFU number. For example, an investigator who has access to data files of croaker and brown shrimp may wish to examine the catch-per-unit of effort of each at the same time to see if there may be a predator-prey relationship.

A third advantage is that retrieving and processing condensed data is cheaper than retrieving and processing raw data.

To illustrate the utility of SDC, we will use an example from our own file data. The menhaden biostatistical master file (2,138 records) and the catch-record master file (575 records) for Beaufort, N.C., were condensed into a single SDC file of 38 PFU records; a hydrographic master file (8,217 records) of temperatures at Beaufort inlet was condensed to an SDC file of 52 PFU records (Figure 1). To avoid unweighted means when PFU values were combined for SFU, TFU, or annual summaries, sums of length, weights, and numbers of fish sampled were used, rather than means. Mean temperatures -- computed by summing the highest and lowest readings per tidal cycle and dividing by the number of tidal cycles -- were used, however, in the hydrographic SDC file.

The SDC file was used to generate output information shown in examples A and B (Figure 1). In A, a graph was constructed showing mean temperature and total catch for each SFU. In B, a table was constructed showing the weight of the catch and the number of samples of the catch for each SFU. For manipulating SDC data we used a computer package -- Statistical Analysis System (Barr et al. 1976). For other types of data other packages might be appropriate.

To summarize, we believe the SDC system is a unique, useful, and relatively inexpensive method of cataloging biological and environmental data for rapid retrieval and processing. Storage and retrieval costs are less because there is one record, rather than multiple records, per PFU. Because data are partially summarized, files on separate but related subjects can be easily merged for a wide variety of comparisons.

# LITERATURE CITED

Barr, A.J., J.H. Goodnight, J.P. Sall, and J.T. Helwig. 1976. A user's guide to SAS-76. SAS Institute, Inc., Raleigh, N.C. 329 p.

Cohen, L.J. 1976. Data base management systems. Q.E.D. Information Sciences, Inc., Wellesley, Mass. 432 p.

Schneidewind, N.F. 1972. Information systems and data requirements: coastal development planning. Pages 221-244 in J.F.P. Brahtz, ed. Coastal zone management: multiple use with conservation. John Wiley and Sons, Inc. New York.

Table 1. Inclusive Dates for Primary (PFU), Secondary (SFU), and Tertiary Fishing Units (TFU).

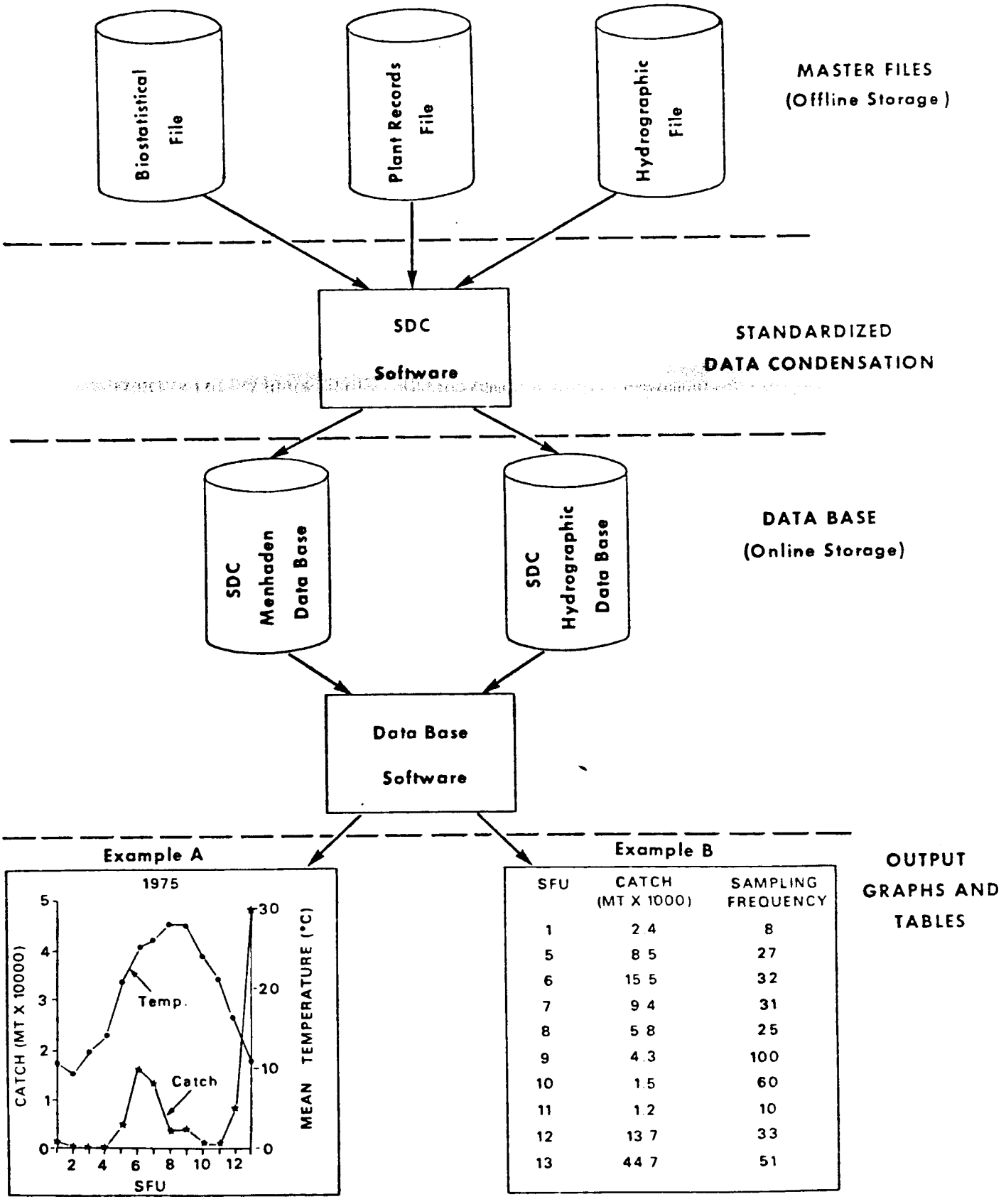
| PFU No. | SFU No. | TFU No. | DATES |
|---|---|---|---|
| 1 | 1 | 1 | 1/1-1/7 |
| 2 | 1 | 1 | 1/8-1/14 |
| 3 | 1 | 1 | 1/15-1/21 |
| 4 | 1 | 1 | 1/22-1/28 |
| 5 | 2 | 1 | 1/29-2/4 |
| 6 | 2 | 1 | 2/5-2/11 |
| 7 | 2 | 1 | 2/12-2/18 |
| 8 | 2 | 1 | 2/19-2/25 |
| 9* | 3 | 1 | 2/26-3/5* |
| 10 | 3 | 1 | 3/6-3/12 |
| 11 | 3 | 1 | 3/13-3/19 |
| 12 | 3 | 1 | 3/20-3/26 |
| 13 | 4 | 1 | 3/27-4/2 |
| 14 | 4 | 2 | 4/3-4/9 |
| 15 | 4 | 2 | 4/10-4/16 |
| 16 | 4 | 2 | 4/17-4/23 |
| 17 | 5 | 2 | 4/24-4/30 |
| 18 | 5 | 2 | 5/1-5/7 |
| 19 | 5 | 2 | 5/8-5/14 |
| 20 | 5 | 2 | 5/15-5/21 |
| 21 | 6 | 2 | 5/22-5/28 |
| 22 | 6 | 2 | 5/29-6/4 |
| 23 | 6 | 2 | 6/5-6/11 |
| 24 | 6 | 2 | 6/12-6/18 |
| 25 | 7 | 2 | 6/19-6/25 |
| 26 | 7 | 2 | 6/26-7/2 |
| 27 | 7 | 3 | 7/3-7/9 |
| 28 | 7 | 3 | 7/10-7/16 |

Table 1. Con't

| PFU No. | SFU No. | TFU No. | DATES |
|---|---|---|---|
| 29 | 8 | 3 | 7/17-7/23 |
| 30 | 8 | 3 | 7/24-7/30 |
| 31 | 8 | 3 | 7/31-8/6 |
| 32 | 8 | 3 | 8/7-8/13 |
| 33 | 9 | 3 | 8/14-8/20 |
| 34 | 9 | 3 | 8/21-8/27 |
| 35 | 9 | 3 | 8/28-9/3 |
| 36 | 9 | 3 | 9/4-9/10 |
| 37 | 10 | 3 | 9/11-9/17 |
| 38 | 10 | 3 | 9/18-9/24 |
| 39 | 10 | 3 | 9/25-10/1 |
| 40 | 10 | 4 | 10/2-10/8 |
| 41 | 11 | 4 | 10/9-10/15 |
| 42 | 11 | 4 | 10/16-10/22 |
| 43 | 11 | 4 | 10/23-10/29 |
| 44 | 11 | 4 | 10/30-11/5 |
| 45 | 12 | 4 | 11/6-11/12 |
| 46 | 12 | 4 | 11/13-11/19 |
| 47 | 12 | 4 | 11/20-11/26 |
| 48 | 12 | 4 | 11/27-12/3 |
| 49 | 13 | 4 | 12/4-12/10 |
| 50 | 13 | 4 | 12/11-12/17 |
| 51 | 13 | 4 | 12/18-12/24 |
| 52 | 13 | 4 | 12/25-12/31 |

* PFU - 9 contains 8 days in regular years and 9 days in leap years.

Figure 1. Flowchart example using SDC to standardize and condense raw data for analysis and report generation




| SFU | CATCH (MT X 1000) | SAMPLING FREQUENCY |
|---|---|---|
| 1 | 2.4 | 8 |
| 5 | 8.5 | 27 |
| 6 | 15.5 | 32 |
| 7 | 9.4 | 31 |
| 8 | 5.8 | 25 |
| 9 | 4.3 | 100 |
| 10 | 1.5 | 60 |
| 11 | 1.2 | 10 |
| 12 | 13.7 | 33 |
| 13 | 44.7 | 51 |