Techniques for Handling Large Pond and Farm Datasets

MARK PREIN
Institut für Meereskunde
an der Universität Kiel
Düsternbrooker Weg 20
D-2300 Kiel 1
Federal Republic of Germany

ANA MILSTEIN
Agricultural Research Organization
Fish and Aquaculture Research Station
30820 Dor, M.P. Hof Hacarmel
Israel

Introduction

Fish culture is a complex system in which biological, environmental and economic factors are involved. Some of these factors are under the control of fish farmers and some are not. Those which are under control include stocking, water exchange and aeration, feeding, fertilization and harvesting. Stocking parameters include date, density, size and species stocked. The stocking decisions affect the entire subsequent operation. With respect to fertilizers and feeds, the farmer must determine the rate, quality and timing. Harvesting can be done on a partial and selective basis during the season in which the farmer must select what to harvest and the timing. Alternatively, a complete harvest can be made and then the ponds can be restocked during the season. These control variables interactively affect biological and economic performance. The situation is further complicated by those variables for which the farmer has only partial or no control, such as length of growing season, weather conditions, fluctuations in prices and market conditions.

Methods commonly used for decision-making are based on general knowledge, experience and intuition of farmers, supported by work carried out at research stations on specific problems which arise during the development of the aquaculture industry. Recent advanced approaches in data analysis allow deeper insights into the functioning of highly complex systems with large numbers of variables. Multivariate statistical methods are the prime choice, since these are powerful enough to deal with the large amount of uncontrolled data and variables involved in aquaculture research and production, and can uncover their complex interactions[1].

The application of these techniques requires vast amounts of data, which are continually produced and accumulated in aquaculture research stations and fish farms. Regrettably enough, up to now this tremendous wealth of original data points is used only once, and its potential information content as a set and as a reference background is neglected.

For the handling of vast amounts of data, methods must be devised to compile all data in a form facilitating further processing, which requires the use of a computer. This tool is nowadays increasingly available and its potential is underused in fish farms. Commonly, a database system must be designed[2], consisting of a series of files of standardized format organized in such a way as to allow easy retrieval of information according to given criteria. In the process of building the database several difficulties may arise.

From research projects presently in progress[3], two examples are described, where vast amounts of data are being assembled in appropriate formats. Each example has its own problems requiring case-specific solutions. In both instances different types of multivariate data analysis will be performed on the compiled datasets, to identify and quantify the effect of key variables governing growth and production of fish in manured pond systems culturing tilapia and carp.

First Example: ICLARM/CLSU Experiments in the Philippines

A four-year research project was conducted at the Central Luzon State University (CLSU) in the Philippines as a joint project between ICLARM and CLSU to develop integrated animal-fish farming systems under tropical conditions.

Numerous experiments were conducted, investigating the growth performance of Nile tilapia *(Oreochromis niloticus)*, common carp *(Cyprinus carpio)* and snakehead *(Channa striata)* in ponds receiving pig, duck or chicken manure as the only nutrient input. The system was designed for application by small-scale farmers. Parameters were varied and measured, including fish size and treatments (pond size, fish, livestock or fowl density); also recorded were meteorological and water quality data. These were assembled in matrix form and partly published in the experimental report[4]. Most of the relevant data was made available on magnetic tape since it had been previously entered on a mainframe computer.

For the presently ongoing project this dataset is being vastly edited. All variables are checked for input errors, since these may cause wrong results in the final

analysis. Variables are plotted over time and several other variables to detect outliers. These points are then checked for plausibility. Printouts of data are searched for missing values. Since only complete cases can be considered in the intended regression analysis, a single missing value in a case would cause the information content of all other variables to be lost. To avoid this, variable-specific missing value treatments are applied[5]. The size of the present ASCII file is approximately 350K and is now stored on disks in PC format.

The original raw data sheets were retrieved and reorganized so that previously omitted but important variables can be added to the data and be included in the final analysis. Unfortunately, some of the raw data record sheets were lost. Up to now the editing of this "historic" dataset has taken several months and is in progress.

Second Example: Historic Production Data of Farms in Israel

Aquaculture in Israel consists of a polyculture of common carp *(Cyprinus carpio)*, tilapia hybrids *(Oreochromis aureus x O. niloticus)* and silver carp *(Hypophthalmichthys molitrix)*, with varying proportions of grey mullet (*Mugil* sp.), grass carp *(Ctenopharyngodon idella)* and bighead carp *(Aristichthys nobilis)*. Management practices vary between regions, farms and individual ponds. Ponds are fertilized with manure and aerated according to oxygen conditions. Fish are fed with various types of pellets with regular sampling of fish sizes. Data are recorded by the farmers on charts for each pond individually.

The aim of the project was to compile a database from commercial production origin covering the years 1980-1986. Sources are approximately 30 farms with a total of over 230 ponds. Although the farmers were very cooperative and willingly supplied their record charts, problems still existed:

- In case of missing values or partial information kept in other locations on the farm (such as applied feed and manure amounts), farmers were reluctant to invest further (often considerable) time to complete the records.
- Since their recordkeeping is designed for commercial management purposes and not for scientific data analysis, not all farms had data of adequate quality for our objectives. Therefore stringent entry criteria for data selection had to be defined prior to data selection and entry.
- Since the farms practically do not record water quality or meteorological data, the latter had to be acquired from the meteorological services. This required definition of variables, their units, geographical regions, years, data purchase, reading from magnetic tape (ASCII tables, integer values) onto disk and subsequent editing.

Central to all data handling procedures is the use of a spreadsheet program on a personal computer for ease, flexibility and standardization. Data were entered into many small files of standard format. These were merged to larger files for analysis and can even be transferred to mainframe computers if necessary. A very important and time saving procedure proved to be the immediate checking of errors after each step:

- The record cards were only short-time loans from the farmers. Photocopies of the record cards would have been useless since the farmers note data for different fish species in different pen colors in the same columns!
- The newly entered data had to be checked against the original cards for typing errors.
- The consistency of all data sequences had to be checked for errors made by the farmers.
- Since it required whole-day trips to visit the respective farms the moment of returning the loaned record cards had to be used for interviews with the farmers in case of unclear or illegible records.

Here too, the process of data entry and editing has taken several months and is continuing.

Conclusions

Both examples are "historical" approaches since the people compiling and analyzing the data are different from those who generated them. This causes various problems up to the point where questions about details of circumstances under which certain data points were created must remain unanswered. After defining the criteria of the intended analysis and designing the required data and their formats, the difficulty of assembling this dataset from various sources arises. This can be a time-consuming task of many months, depending on locations and sources, also requiring extra personnel for data entry. Generally it is advantageous to enter raw data into small, standardized spreadsheets which also facilitate the frequent checks at all stages of the data entry process. This also applies to currently running experiments where the collected data should be immediately (at best daily!) entered into the database. If possible, cases should always be complete.

The compilation of a database is useful also for normal bivariate comparisons between ponds or years, even if advanced multivariate analyses are not intended. The recommended hardware requirements are: PC (IBM-compatible) with 20MB hard disk, 640K RAM, math coprocessor, and printer.

IMPORTANT: all files should be stored at least in duplicate on backup disks in a safe place! All compiled raw data should be made available to the scientific community in the form of data tables in appendixes of reporting publications and/or on diskettes. By not doing so, the whole idea of reproducibility of scientific results becomes awkward. There are always chances that as analytical methods improve, new insights may be gained from already available data. The data and results of the present project will be published in an ICLARM Technical Series.

---

1 Milstein, A., G. Hulata and G.W. Wohlfarth. 1988. Canonical correlation analysis of relationships between management inputs and fish growth and yields in polyculture. Aquaculture and Fisheries Management 19:13-24; Pauly, D. and K.D. Hopkins. 1983. A method for the analysis of pond growth experiments. ICLARM Newsletter 6:10-12.

2 See also: Hopkins, K.D., J.E. Lannan, J.R. Bowman. 1988. Managing a database for pond research data - the CRSP experience. Aquabyte 1(1):3-4.

3 Optimal Management of Aquaculture Pond Systems in Developing Countries under the auspices of the German-Israeli Fund for Agricultural Research in Third World Countries.

4 Hopkins, K.D. and E.M. Cruz. 1982. The ICLARM-CLSU integrated animal-fish farming project: final report. ICLARM Technical Reports 5, 96 p.

5 Prein, M. 1985. The influence of environmental factors on fish production in tropical ponds investigated with multiple regression and path analysis. M.S. Thesis, Kiel University, 91 p.