Developing an Aggregated Survey/Macro-economic Database for Statistical and Graphical Social Science Applications

Michael B. Zartman and Christopher J. Gordon
Institute for Social Research
The University of Michigan

The Survey Research Center at The University of Michigan has routinely conducted surveys of consumer attitudes since 1946. The May 1981 survey is the 152nd in this series which provides regular assessments of consumer attitudes and expectations. The surveys are designed to explore why changes in consumer attitudes and expectations occur, and how these changes influence consumer spending and saving decisions. A major research objective of the project is to use this collected data to evaluate economic trends and prospects.

Each survey contains "standard" questions asked at regular intervals, many of which have been included from the project's inception. The aggregated results of these surveys provide a wealth of time-series data with the potential to be an important factor in forecasting consumer behavior. The "standard" questions themselves can be disseminated into approximately 190 separate data series (including index transformations). When "nonstandard" (or non-core) questions are included, this total jumps considerably. With such a large number of data variables, many different areas of analysis are available to be researched. When the many macro-economic data series (e.g., Federal Reserve, Census, or Retail Sales data) are added to this compilation, the data management problems increase. The research results which could be achieved, then, are directly related to the development of a flexible method of data storage and retrieval.

Data Storage Systems

Until fairly recently, the Surveys of Consumer Attitudes project staff entered data by hand into various statistical programs. This was very time consuming and limited the numbers and types of analysis which could be performed. Since the data was not stored, it had to be hand entered each time, causing discrepancies from analysis to analysis (with copying errors). Finally, the need became clear to develop a data storage system which could be used to promote consistency between analysis runs. The major parameters which concerned us were accuracy in storing the data, expediency in updating and use, analysis flexibility, and quality reporting capabilities.

©1981 ACM 0-89791-056-7/81/0500/0045 $00.75

Accuracy

The prevention of data errors must begin at the critical step of loading the database. Even if the data is typed in by hand, the non-systematic error introduced when the data is typed in more than once is eliminated. Of course, the best way to limit human error is to limit human interaction. If the data can be computer generated (which is the case for our survey data), then errors in loading could be eliminated.

The problem of errors introduced by hand editing is also present when the database is updated for new observations. The data storage system should be able to allow updates and revisions in an efficient manner. Another plus would be the ability to perform computer accuracy checks on the data (if a systematic way of doing this is available).

Timeliness

Another concern of our project is the speed at which we can turnover our results. This expediency is affected at several points: the process of updating the data series, the task of analysis, and the final stage of reporting the results. If the data must be updated by hand, a great deal of time is lost at this point. Considering the large number of variables concerned, this is the step at which an effort should be made to cut processing time--hopefully, with computing techniques.

Likewise, some processing time could be saved in the areas of analysis and reporting if "standard" regressions or tables are to be produced after each survey. In this case, computer setup files could be maintained which perform the same types of analysis each time after the addition of new data.

Analysis Flexibility

The desired system for data storage should also include sufficient data transformation and analysis capabilities to perform the statistical functions necessary for social science research. Many times, econometric transformations (e.g., moving averages or seasonal adjustments) are desirable for advanced analyses. The storage system should also be able to save these new data series for other analysis purposes. Regression capabilities are equally as important in econometric research. The greater the number of types of regression tasks (e.g., OLS, GLS, two-stage, etc.) available to the researcher, the more complex his analytical design can become, with a greater ability to test the validity of the data.

Another desirable feature in the analysis area of a system is some method of storing regression equations. This is essential in developing simulation models needed to forecast future economic trends. The ability to store these equations would greatly speed up the regression tasks after each new survey.

Reporting Capabilities

The reporting function is the final step in our system design. This is the point at which accuracy in methods, expediency of turnover, and analysis results come to fruition, since without quality reporting mechanisms, the research results would be dated before they could be published in a professional manner.

Tables and charts are an important part of reporting analysis results. They provide visual support for trend studies, and can simplify concepts for quick reference. Producing these as "camera ready" output then, is a vital task for a research system.

Former SCA Time-Series Methods

The first attempt at developing a system to handle our data storage and analysis needs involved MIDAS (Michigan Interactive Data Analysis and Storage) which was created at The University of Michigan. MIDAS allows both data storage and analysis capabilities--essential elements in utilizing time-series data. This statistical package was convenient for our project since it was "in-house" designed and all members of our staff were familiar with its use.

However, even though it was designed for time-series analysis, MIDAS has many weaknesses. (It was not designed to handle the massive amounts of data which our project generates.) It does allow the user to read data in with standard FORTRAN formats, and thus is quite efficient in loading data.

Unfortunately, it is quite inflexible when it comes to updating data series. It does not simply allow additional cases to be added, but forces a re-loading of the entire database. Our processing at that time also relied upon hand coding for all data vectors (including the survey results). The project staff then began a search for a more efficient system of handling our data with the major parameter goals more firmly in mind.

New SCA Time-Series Methods

The Surveys of Consumer Attitudes project staff considered several alternatives to the MIDAS system including developing an entirely new statistical package, however, the cost of such an effort was deemed prohibitive. TROLL (Time-shared Reactive On-line Laboratory), another econometric package already available to University users, seemed to provide sufficient capabilities to meet our needs. TROLL was developed at MIT in conjunction with the National Bureau of Economic Research and was designed with large time-series databases in mind.

Data can be loaded into TROLL "archives" (data storage segments) easily by using a built-in database facility, where each data series is stored as a separate vector file of varying length. Input does not have to be a rectangular matrix (saving disk space and CPU time). TROLL also provides for a user programmable FORTRAN subroutine, allowing loading and updating of data vectors in archives from nearly any type of database design. The "archive" organization of data storage allows users on separate computer accounts to access each others' data selectively (when permission has been assigned).

The NBER maintains TROLL database "archives" of macro-economic data--reducing the SCA's reliance upon entering macro-economic data by hand. The survey data is updated from a data depository file filled with aggregated results of full sample data, as well as subgroup data (e.g., income levels). Previously too numerous to be used in our time-series processing, it can now be entered quickly, accurately, and automatically by the TROLL subroutines.

TROLL offers many analysis techniques as well as data transformation capabilities. Some regression tasks available are OLS, GLS with first and second-order autocorrelation correction, two-stage least squares, and three-stage least squares. Residuals and predicted values can also be saved as permanent data vectors. Data transformation capabilities include moving averages, seasonal adjustments, exponential smoothings, ratios, etc.--all of which can be saved as permanent vector files.

Printing and plotting mechanisms are provided by TROLL for quick analysis support. However, this type of output is not in publishable form. The user programmable, FORTRAN database subroutine offers an efficient method of solving this problem. By writing data to system resident chart files, and then processing these through a graphics routine into a machine readable form, "camera ready" charts can be created quickly and automatically. Furthermore TROLL features a report-writing task which can be used to produce publishable tables.

Conclusions

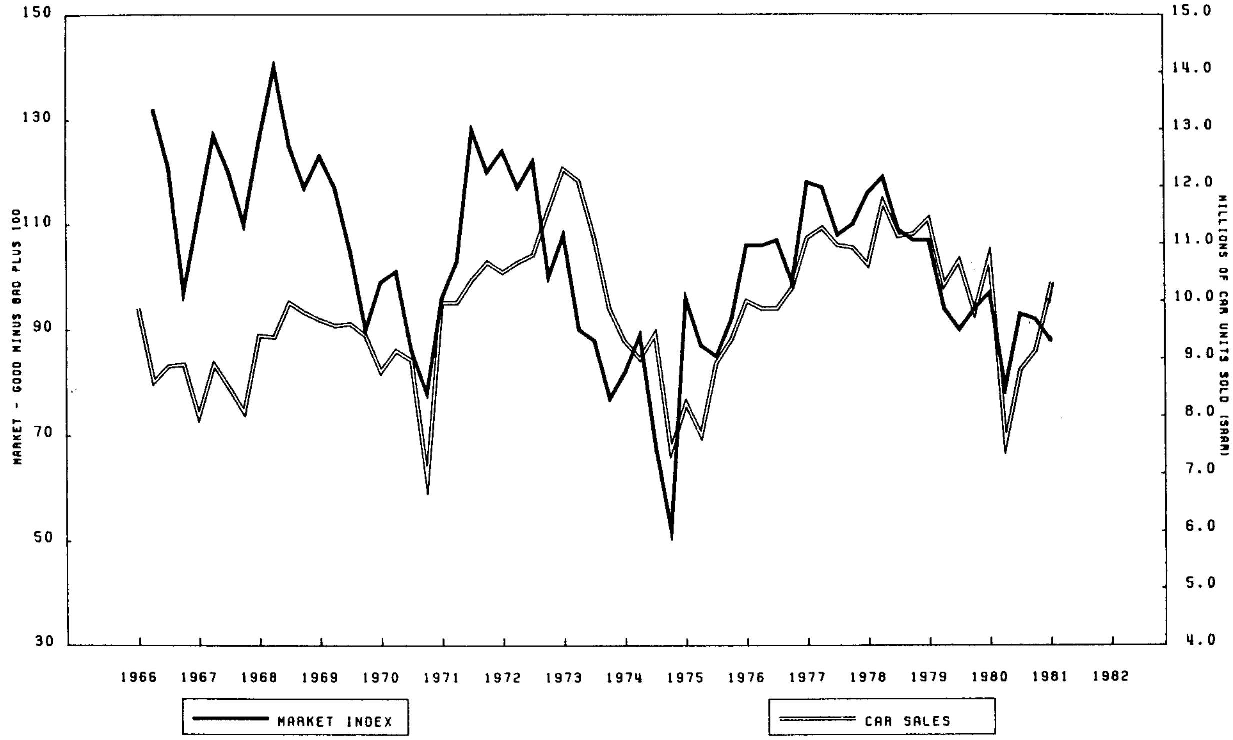
TROLL is quite adaptable to our project's needs for very large data storage capacity, quick turnover, little human interaction, and reporting capabilities. It seems to have been "made to order" for our particular research purposes, although other research groups involved in time-series applications might consider other packages which emphasize different aspects more useful.

Our major research objective, as stated earlier, is to eventually evaluate economic trends and prospects by studying these aggregated survey/macro-economic data. TROLL includes a modelling language and structure which allows regression equations to be stored and edited. A simulation task also is provided to aid in forecasting.

These advantages will not only be utilized by SCA staff, but will eventually be shared by our project's sponsors. Since TROLL allows a user to permit data to be read by others, eventually all sponsors will have access to our survey data via TROLL archives, allowing them to perform their own research tasks.

EXAMPLES

BUYING CONDITIONS FOR CARS AND TOTAL UNIT CAR SALES



	Jan- Mar. 1978	Apr- June 1978	Jul- Sept 1978	Oct- Dec. 1978	Jan- Mar. 1979	Apr- June 1979	Jul- Sept 1979	Oct- Dec. 1979	Jan- Mar. 1980	Apr- June 1980	Jul- Sept 1980	Oct- Dec. 1980	Jan- Mar. 1981
A. BUYING CONDITIONS FOR CARS[*]													
GOOD TIME TO BUY	50%	52%	46%	44%	46%	41%	38%	42%	44%	34%	40%	40%	38%
UNCERTAIN; DEPENDS	17	16	17	19	15	12	14	10	9	11	13	13	12
BAD TIME TO BUY	33	32	37	37	39	47	48	48	47	55	47	47	50
TOTAL	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
B. TOTAL UNIT CAR SALES													
MILLIONS OF UNITS (SEASONALLY ADJUSTED ANNUAL RATES)	10.6	11.7	11.1	11.2	11.4	10.3	10.7	9.8	10.8	7.5	8.8	9.1	10.2

[*] The question was: "Speaking now of the automobile market--do you think the next 12 months or so will be a good time or a bad time to buy a car?"