Abstract

Bureau of Labor Statistics produces the Consumer Price Index (CPI), a major economic indicator, on a monthly basis. The systems and associated subsystems that support its calculation are being revised, with major components written using SAS(R) software. Experiences in using SAS software on a large system that processes high volume data are discussed. The paper addresses the problems encountered, along with a discussion of how they were addressed using the SAS System.

Introduction

The Consumer Price Index (CPI) is perhaps the most misunderstood statistic produced by the Bureau of Labor Statistics (BLS). Mistaken for a measure of inflation or change in the cost of living, it measures price changes for a fixed market basket of goods and services. Changes in consumption patterns are reflected only in periodic revisions, and a number of cost of living components are not included (i.e. taxes other than sales tax, non-cash fringe benefits, life style changes, or government supplied services). Still, it remains the best indicator of inflation on which to base economic policy and measure its success or failure. The CPI is used as a deflator of other economic series such as the GNP, retail sales and other important indicators to provide estimates in inflation-free dollars. Used as a benchmark in escalator clauses, it controls wage increases for more than 8.5 million workers covered by union contracts. Additionally, it determines cost of living adjustments for social security recipients, changes in food stamp compensation and school lunch subsidies, and increases in the official poverty threshold. It is also used to adjust rental fees and child support payments. The CPI directly affects the lives of over fifty million workers, expanded the CPI to include new commodities, used modified weights to describe the relative importance of commodities in aggregating information to a single index, and adopted the use of imputation, that is, the estimation of weights for uncollected items. During World War II, temporary adjustments in weights were made to estimate the impact of rationing and war-time shortages. The 1953 revision, began in 1949, was disrupted by the unexpected Korean conflict and the sharp price increases on a number of diverse items. Changes in this revision included modifications in weights to reflect new consumer expenditure patterns, and the addition of restaurant meals and owned homes as well as other items to be priced for the first time. The 1964 revision included persons living alone and changed income limitations to be included in the survey, but was primarily a badly needed update of weights and priced items to reflect some of the most substantial life style changes in American history. These included a large real rise in disposable income, credit availability and shifts in geographic population distribution. The most recent 1978 revision was concerned with updating weights used with various categories (food, clothes, etc.), items to be priced in the CPI, and the sample of retail stores from which to collect prices. As BLS began to take advantage of the power provided by computers, modernized conceptual and statistical solutions could be incorporated into the CPI. A new index was published covering 80 percent of all urban workers in addition to the traditional wage earners index; separately published indexes were increased from 24 to 20 cities. The number of items priced monthly was increased; bi-monthly pricing replaced quarterly pricing for items surveyed at that frequency; and retail outlets surveyed were more representative of those frequented by consumers. The current revision, in progress, again updates weights for expenditure categories, items to be priced and locations from which to collect prices. It incorporates major survey changes developed after the 1978 revision, and reflects major advancements in a new methodology, rental equivalence, to estimate the price and its subsequent change for cost of shelter and associated amenities such as appliances for homeowners. It is the role of automated data processing and management and the use of SAS in a production environment within the context of the current revision that is the focus of this paper.

Components of the CPI - An Overview

There are a number of questions to answer in the production of a single index to measure the change in price of a fixed market basket of goods and services; what do people buy, where do they buy it, and how much do they pay. To reflect changes in consumer expenditure patterns, this market basket must be updated periodically. The more frequently these changes occur, the more often the market basket covered in the CPI should change. The Consumer
Expenditure Survey (CES) is intended to address the question: “What goods and services do people spend their money on?” The Consumer Point of Purchase Survey (CPOPS), on the other hand, addresses the question, “where do people purchase these goods and services?” These two survey instruments have been instituted on an ongoing basis since the last revision of the CPI. This allows new revisions to weights and items to be priced to be realized in a more timely and cost effective manner. Further, published indexes in past revisions have lapsed the collection of data by years, while the ongoing implementation of these surveys ameliorates this problem.

The information from CES, used to determine the weights, or relative importance of different types of expenditures as compared with information from CPOPS to collect information on prices in 85 cities. The collection of price data is facilitated by two surveys, Commodities and Services (CAS) and Housing. Under contract to BLS, Bureau of Census field personnel are provided with survey forms and instructions to collect retail price information from selected stores and housing prices from individual sampled housing units. Retail prices to be collected vary in number due to seasonal items, sample design rotation, product discontinuation, closure of retail outlets, and elimination of goods from a sampled retail outlet. Estimates provided by BLS staff indicate that, typically, between 110,000 and 111,000 individual prices will be collected each month in the current revision. The prices collected can be broken into seven categories: food and beverages, housing, clothing, transportation, medical care, entertainment and other goods and services. Some priced items, such as train fares, some types of insurance and other intractable items are collected by telephone from BLS headquarters rather than by survey interviewers. Additionally, approximately 60,000 housing units (10,000 units per month over each 6-month period) are interviewed to determine shelter costs and accommodations such as number of rooms, air conditioning equipment, utility costs and a host of other items totalling as many as 600 or more individual pieces of information. Through a complex series of editing, imputation and estimation methodologies, price changes from the previous month are computed and aggregated to produce a single all-items index at the national level for two populations: wage earners 600 or more individual pieces of information. and a host of other items totaling as many as 600 or more individual pieces of information. Estimates provided by BLS staff indicate that, typically, between 110,000 and 111,000 individual prices will be collected each month in the current revision. The prices collected can be broken into seven categories: food and beverages, housing, clothing, transportation, medical care, entertainment and other goods and services. Some priced items, such as train fares, some types of insurance and other intractable items are collected by telephone from BLS headquarters rather than by survey interviewers. Additionally, approximately 60,000 housing units (10,000 units per month over each 6-month period) are interviewed to determine shelter costs and accommodations such as number of rooms, air conditioning equipment, utility costs and a host of other items totalling as many as 600 or more individual pieces of information. Through a complex series of editing, imputation and estimation methodologies, price changes from the previous month are computed and aggregated to produce a single all-items index at the national level for two populations: wage earners and clerical workers and all urban workers. Indexes for selected cities are also published.

What’s Not in SAS Software

In the production of the CPI, a number of tasks are performed: data collection, data entry, data editing, data storage, and other data processing. While the majority of required data processing is done using SAS software, not all tasks are performed using this tool.

Once surveys are completed, the forms are transmitted to headquarters. Data entry and preliminary data editing are performed using micro computers with software written in PASCAL. The use of micro computers is enhanced by a micro computer configuration linked to a Britton Lee machine which allows operation typical of a relational data base. Thus, previous price data can be extracted from the mainframe data base to make comparisons and reload any changes. Data entry via SAS/FSP(R) and TSO uses significantly more resources than data entry via micros. Hence, the latter solution is preferred.

Since the Britton Lee has a limited data storage capacity, a tape is prepared as percentage utilization approaches its maximum. In the standard tape format for a SAS data set, the information is uploaded through an interface into the mainframe data base. RAPID, a relational data base product from Statistics Canada, is used to store the edited survey data. The relational nature of the data base is ideally suited to the organization of the survey forms and associated collected data. SAS data sets are rectangular in nature and storage of records which include missing values uses excessive disk space, thus the SAS System is not an efficient data storage tool in this case. Further, the UPDATE statement is slow and inefficient when only a small percentage of records and fields on records in a large file are updated frequently as is the case in CPI data processing. The relational data base permits access to a wide number of users, both production systems and researchers, without a processing language restriction and provides powerful capabilities typical of a relational data base. To assist users, BLS developed and maintains an interface to download data selectively from RAPID relations to SAS data sets. To provide even greater flexibility, additional interfaces have been under contract for the Bureau which upload the contents of a SAS data set directly into RAPID relations. These features, data collection, preliminary editing and data storage are the only processing components associated with production of the CPI that are not designed entirely using SAS software.

What’s in SAS Software

Before discussing the role of SAS software in the revision CPI, a review of the approach to systems development at BLS is in order. Using a traditional systems life-cycle approach and structured analysis, the design and implementation of each sub-system consists of several phases: analysis, system specification, system design, program design, prototyping and testing. The first phase, analysis of user requirements can be assisted in a timely fashion with ad-hoc SAS confirm methodological and other assumptions on the part of the user. The second phase, system specification, is a functional description of how the system will operate, from which the basic SAS structures needed can be extracted. Once system specification has been accepted by the user, system and program design can commence. Program design expresses functions in physical terms which can be easily translated into SAS DATA or PROC steps where flows of data between processes become SAS data sets. The data independence of SAS software as a programming language provides significant design flexibility allowing dynamic data-driven solutions to data processing problems. The use of the SAS Macro facility in system design to conditionally compile and execute code provides tremendous power and flexibility and affords more of the advantages of more traditional, procedural languages than does base SAS software alone. These include use of Macros to package code into “main” programs and “subroutines” which can then be conditionally invoked using %IF Macro logic and controlled by parameters in
the form of Macro variables. Prototyping can be done quickly using SAS software to illustrate the features of new systems, demonstrate the feasibility of design concepts, identify design weaknesses, or compare the efficiency of different coding solutions. The use of SAS Macro makes testing of programs much easier through the conditional execution of prints and other reports. While in production, these diagnostics can be switched off with a single parameter. If a problem is encountered or retesting is required following changes or enhancements, the verification tools can be easily reactivated. The use of SAS Macro in system design and development is more thoroughly discussed in "The Use of the Macro Facility in a System Development Environment" (Jeff Phillips, DRL, Inc., SUGI '86 Proceedings). Further, SAS PROCs can assist the user in verifying results and accepting the design and its implementation.

The advantages of using SAS software in combination with structured design extend through the production life of the system or subsystem. In most systems which are run for several years, a number of changes are at least considered, if not enacted. Structured design documents can be consulted for each type of change: minor maintenance, replacement methodologies or enhancements/additions to the system. Changes to the system can be expressed in terms of the design and then easily translated into code. Since the system's code parallels the design document, altered portions of the design correspond to SAS DATA and PROC steps that can be lifted and replaced by the code corresponding to the redesigned portions, without affecting the remaining program modules. Testing of the redesigned modules can be facilitated as described previously using SAS Macro and diagnostic tools. The compatibility of the replacement modules with the original system can be determined through an examination of any new SAS data sets using tools such as PROC CONTENTS to ensure that new or modified SAS data structures correspond to what is expected by the unaffected program modules.

Having examined the nature of system design and development at BLS, and the advantages of using SAS software with such an approach, a review of the types of data processing performed in the CPI using SAS is in order. Included are imputation, data management, calculations, reports, and survey form generation.

SAS data libraries and the data management features of the SAS System are used widely in a number of CPI systems and subsystems. One example was described previously in the process of uploading data from the Britton Line configuration of PCs to the mainframe relational data base. Communication of information between subsystems as SAS data libraries is common, allowing ease of access and subsequent data transformation and contributing to a uniform system architecture. Additionally, these structures and features are used in tracking files which maintain an historical record of a specified duration of intermediate results of data processing in a number of systems. These files are subsequently used in research for improved techniques and methodologies for computation of the CPI and its components.

The design and coding of imputation methodologies using SAS software is widespread in the CPI system and the processing of CES, CPOPS, CAS and Housing survey data. Imputation is the process of filling in missing values, a vital consideration in a survey data processing system where respondents may not know an answer may be unavailable or when an item to be priced has been discontinued. A number of different statistical methodologies are used in a variety of subsystems; hot deck matching, percentage distribution, weighted means and allocation. Some subsystems employ a hierarchy of methods where SAS software is used as an automated decision maker to examine the results of an imputation and select a superior or more appropriate method. One of the most common methods, hot deck matching, has evolved to include a consistent set of SAS software features. Hot deck matching is the imputation of a missing value using other characteristics of the data to select a best match based on the correspondence of these characteristic variables. Typically, four to six variables are used for characteristics; PROC SUMMARY is then used with the CLASS statement to specify these variables and determine the number of records in every combination of values for each possible permutation of variables. The best's' priority of permutations can be redefined by formatting the precedence indicator supplied by PROC SUMMARY for each permutation. The best matches for each imputation are then randomly selected. These and other imputation methodologies are discussed in more detail with examples from the CE survey in the referenced data adjustment paper (Mopsik and Dippo).

Separately from imputation methodologies, a tremendous number of calculations are required in the CPI, as is typical of many production systems. There are, of course, the usual kinds of data transformations found in many production systems; mathematical operations, recoding or reformating, binary operations, string operations, logical operations, and other less common functions (e.g., hexadecimal operations). While these capabilities are available in procedural languages aside from the SAS System, the advantages of SAS software described earlier make it a superior tool to capture the complexities of the estimation methods that employ these various types of calculations. Some examples of these methods include determination of relative importance, weighted aggregation, relative computation, and allocation. The computation of relative importance of purchases, or cost weights, determines which items comprise what portion of expenditures (e.g., gallon of milk versus gallon of gas). Relative importance is also used at the price collection level where an item is priced (e.g., share of buttermilk versus skim milk at a given retail store). The relative importance is used to weight price data and aggregate this information into more general categories (e.g., milk, food, bread, etc. into food). Price movements, chair relatives, are computed at particular levels of aggregation based on these weighted costs. Relatives are a vital component in determining the value of the CPI in each month.

Another estimation method, allocation, is used in a number of CPI systems and subsystems. Allocation involves the disaggregation of reported prices. Examples include apportionment of reported utilities cost performance based on an individual's reported appliances (e.g., gas stove, electric dishwasher, solar heat) or the
apportionment of reported prices, based on research, into new uncollected categories to reflect new expenditure patterns prior to a major revision or survey modification.

Numerous reports are produced in the computation of the CPI. The SAS System is particularly well suited to respond to these requirements of the CPI. Reports range from test prints and other prepackaged formats which can be structured according to user requirements (e.g., PROCs CHART and PLOT) to fully customized report layouts. Through the use of some sophisticated report writing features in the SAS System, data driven reports in customized layouts can be produced where the placement of items on a report page is determined by computed results.

One application of these features is the customized generation of survey forms for the CPI survey. Using a number of data structures which describe what items to price (varying by geography and season), where to price items (varying with sample rotation), and previous price data, forms are dynamically generated. A separate form is produced for each retail outlet with the items printed in the order to be priced with their cost from the previous pricing period which varies by item. The forms are printed in the sequence to be used by field personnel in visiting each retail outlet. This system avoids needless confusion and simplifies field work; data collectors price only those items on the form and note only prices which have changed since the last time they were collected.

Typical Objections to the SAS System in a Production Environment

The role of SAS software and its appropriateness for the computation of the CPI has been discussed as an example of the utility of the SAS System in production environments in general. There are a number of objections frequently raised by system designers and developers to the use of SAS software in a production environment; limitations to sequential data processing, inability to store compiled code, lack of DBMS interfaces, and inefficiency. There are effective arguments to be made against each of these objections.

The belief that the SAS System is capable of sequential processing only is mythical. While there are not limitless data input options, there is some flexibility. Direct access of records in a file can be accomplished through options on the INFILE statement. Further, the POINT option of the SET statement can be used for direct access of any SAS data set, which is a strong argument for maintaining data in SAS data sets and data libraries when practicable.

The inability to store and subsequently execute compiled code in the SAS System may be available in the future. Regardless, there are some advantages to the use of non-compiled code. In bookkeeping terms, the user of a program does not have to be concerned that the object code module used reflects the most recent revisions in source code since they are not separate entities in the production system. Furthermore, with the introduction of SAS Macro, steps can be loaded into Macro variables whose values are subsequently used to control what code is emitted to the compiler using IF logic. Thus, the code to be compiled within a given program module can be data driven and dynamic, varying from run to run. This is incompatible with the concept of using object code which does not vary.

The lack of DBMS interfaces is another typical argument against the use of the SAS System in a production environment in either a primary or complimentary role. There are actually a number of solutions to this problem. First, an exhaustive survey of available products and their suitability should be conducted. There are some DBMS interfaces written and supported by SAS Institute (e.g., for IMS, SYSTEM 2000). Others have been developed by the DBMS vendor itself (e.g., for MODEL 204, INQUIRE, and FOCUS). Other products developed by third parties are also available for sale or lease (e.g., for ADABAS, RAMIS, and TOTAL). If none of the available products satisfies the DBMS requirements and provides the user the necessary capabilities, an interface can be developed if the resources are available and the need justifies the cost. PROC RAPSAS, data extraction, and PROC RAPDAP data base update, built to interface the SAS System and RAPID at BLS are but one of many examples.

The most common objection to the use of the SAS System in a production environment is a perception of inefficiency. Frequently, inefficiencies in SAS programs are the result of design flaws or a lack of understanding of how SAS software works. There are a number of techniques that can be used in different situations to reduce CPU time. This is not the only type of efficiency to be considered in the design, development and maintenance of a system. However, other concerns such as personnel costs and frequency of use must also be considered as well. If a program is not to be used frequently, it is often easier to design and code quickly in SAS software than in other languages, reducing personnel costs. For programs which are executed more frequently and have a longer life span, the advantages of using the SAS System throughout the course of the system life cycle have been discussed previously. It is these larger systems, which run more frequently, where techniques to reduce CPU time and other computer resources consumption are most beneficial.

A number of papers have been written offering approaches or solutions in SAS software to reduce computer resources, many of which are referenced. Some examples of these techniques are listed below:

- Use of PROC SUMMARY to compute counts and selected descriptive statistics over PROCs MEANS or FREQ.
- Use of PROC FORMAT and the PUT function for table look-up to recode values or subset selection of data.
- Use of Macro to generate hardcoded statements rather than using arrays.
- Use the LENGTH statement to minimize needed disk space.
- Use DROP or KEEP as data set options to reduce I/O.
- Place the SET statement within a DO WHILE loop, when BY group processing is not required, to reduce the number of executions of the DATA step.
- Use the formatted form of the INPUT statement.
- Use the trailing 0 sign on the INPUT statement to read only required information in hierarchical files or files with variable length records.

Conclusions

To summarize, the current CPI revision is a large, complex production system for which the SAS System was chosen as a primary software tool. The data independence and non-procedural nature of SAS software provides a natural correspondence between structured analysis and design techniques and SAS coding efforts, making development, testing and maintenance faster and easier. In addition, the SAS Macro language provides all of the advantages of traditional programming languages.

Although there are a number of typical objections to the use of the SAS System in a production environment, some are based on misconceptions, and others can be circumvented using some of the ideas and techniques described above.

SAS and SAS/FSP are registered trademarks of SAS Institute Inc., Cary, NC, USA.

Acknowledgements

The authors would like to express their appreciation to the staff of Consumer Purchases and Consumption Studies, Bureau of Labor Statistics for their assistance in developing this paper.

References


Mopsik, Judith. "Efficient Techniques for Large Data Files," SUGI '83 Proceedings, SAS Institute, Cary, N.C.


Sharlin, Joshua. "Data Reduction and Summarization," SUGI '83 Proceedings, SAS Institute, Cary, N.C.


The authors may be contacted at:

ORI, Inc.
Software Applications and Training Division
122 C Street, N.W., Suite 250
Washington, D.C. 20001
(202) 737-2666

340