

THE EVOLVING ROLE OF THE SAS® SYSTEM IN THE DEVELOPMENT OF THE PRODUCTION SYSTEM FOR THE CONSUMER PRICE INDEX (CPI)

James E. Johnstone, Bureau of Labor Statistics
Stuart Leven, Bureau of Labor Statistics

Abstract

The Bureau of Labor Statistics (BLS) produces a number of economic indicators measuring changes in various market sectors, among them the Consumer Price Index (CPI). This index measures, using a probability sample, the change in price of a fixed market basket of consumer goods and services in specific geographic areas as well as by region and across the U.S. The impact of the index is far reaching. In the 1986 SUGI Proceedings, the role of the SAS® System in the development of the revision of the associated production system was discussed. This revision came on-line in 1987. This paper discusses experiences in five years of production use of the SAS system in this environment, the faults and foibles uncovered, and the role of the SAS system in maintenance in a real-life setting of this large-scale system. It also addresses the role of the SAS system as foreseen in the next revision of the CPI, scheduled to come on-line in 1997, as well as visions for the future in both production and ancillary research for ongoing improvements to the CPI.

INTRODUCTION

Purpose and Organization

The purpose of this paper is to demonstrate that the decision in the early 1980s to use the SAS system in this large-scale survey production environment was sound as demonstrated by the level of success within the CPI program. In addition, the paper demonstrates that the SAS system has played and will continue to play a significant role in software development for the production system and research that fuels changing requirements to improve the index.

The remainder of the introduction is devoted to a brief description of the index, its economic impact, and its historical background. An overview of the CPI surveys is provided, followed by a brief discussion of the complexity of the CPI. The most recent revision is then described in terms of motivation for the conversion to SAS software and the impact of structured systems development. The 1987 revision is then discussed in terms of platforms, processing tasks, and the roles in which the SAS system serves followed by a discussion of changes since the revision software came on-line. Next, the perceptions of how well SAS software has performed in our environment of ongoing system changes and enhancements are enumerated. We then go on to relate visions for the future in

terms of the role of the SAS system in economic and statistical research as well as prototypes and ongoing development. The future visions section concludes with an exploration of the role of the SAS system in the upcoming 1997 revision and beyond.

The Index as a Measure

The Consumer Price Index is a modified Laspeyres index representing a ratio of costs of purchasing a set of items of constant quality and quantity in two different time periods. This fixed set of items is referred to as the consumption market basket. The ratio of these costs from month to month is called a price relative and is used to inflate the index values from a previous time period to the current. As such, the CPI is an indicator of the cost of living.

Changes in consumption patterns are reflected periodically in the CPI, rather than on a continuous basis. A number of cost of living components are outside the scope of what the CPI is intended to measure. These include taxes other than sales tax, non-cash fringe benefits, and government supplied services. The role of the surveys which fuel the index and the difficulties in reaching a constant quality/quantity criterion are discussed in more detail below. The reader is referred to the BLS Handbook of Methods (1988) for further details concerning the index as a measure, computational formulae, and statistical concepts employed in various methodologies.

Economic Impact

The CPI is widely used as an economic indicator in a variety of economic models. It is used as a deflator in the Personal Consumption component of the Gross Domestic Product (the recent replacement for the Gross National Product), retail sales and other important economic measures to provide estimates in constant dollars. In addition, the index has a direct impact on the livelihood of more than half of the U.S. population. The index is used within federal programs to adjust cost of living increases in social security, school lunch subsidy programs, food stamp allocation and other welfare benefits, and as an inflator for the official poverty threshold. Within municipal programs, it is widely used as an inflator or limit for rental payment increases, child support and other court stipulated payments, as well as other distributions from municipal funds. In addition, it is used within many wage earner contracts to determine increases in pay and other forms of compensation. Errors in the CPI at the national level as small as 0.1 percent can

result in misdirection of hundreds of millions of dollars within the national economy.

Historical Background

The measurement of price change began in the late nineteenth century with special studies commissioned by Congress. A cost of living index was first published in 32 cities in 1919 to adjust shipbuilders' wages. In 1921, regular periodic publication began and moved to a quarterly basis by 1935. Over the following decades, the CPI was improved and expanded many times. With each revision, the items to be priced and their associated weights used in estimation were updated. For example, expansion of coverage included restaurant meals and owned homes in 1953.

The 1978 revision updated weights, the items to be priced and the sample of outlets from which to collect price information. It represented the first time that the pricing surveys were based on a full multi-stage probability design for sampling (Dippo and Jacobs, 1983). The advent of high speed computers allowed the scope of coverage and the frequency of pricing to increase substantially. Improved data processing capabilities allowed for the introduction of more rigorous conceptual and statistical solutions for a variety of estimation and computation problems.

The capabilities of high speed computers were even more evident in the 1987 revision. This revision provided for the usual updates in weights, items to be priced and locations in which to price them. It substantially modified the implementation of rental equivalency, a concept first introduced into the CPI in 1983. Rental equivalency is an attempt to measure the change in costs of shelter and associated amenities within owned homes through exclusion of such items as capital appreciation and investment. The revision introduced much more sophisticated imputation techniques (the estimation of values for missing items based on other information within the survey), which are data-driven in their application. The revision also represented a complete conversion of mainframe systems to the SAS system.

THE CPI SURVEYS

Surveys and Their Objectives

To generate the CPI, one needs the answers to a number of questions. These fundamental questions comprise the objectives of each of several surveys and can be encapsulated simply: what do people buy, where do they buy it and how much do they spend.

First, one must know what goods and services people purchase at the retail level. This is the market basket which reflects consumer expenditure patterns. As these patterns change, the CPI market basket must also change. To address this issue, the Consumer Expenditure Survey examines purchasing patterns on

an ongoing basis. Essentially, this survey measures, at a very detailed level, how much the sampled households spend on a wide variety of consumer goods and services. Through this survey, the items that comprise the market basket and their relative importance (reflected as weights) are determined.

Once the consumption market basket is determined, information on where people purchase these goods and services at the retail level is required. This is the essential objective of the Continuing Point of Purchase Survey. The results of this survey, also collected on an ongoing basis, together with the results of the Consumer Expenditure Survey are used to determine on what and from where to collect prices in a wide variety of urban areas.

The monthly pricing for the CPI itself comes from the Commodities and Services (C&S) and Housing Surveys which periodically collect prices and supporting information. The essential objective of the C&S Survey is to collect price quotations (termed quotes) and related information for retail goods and services from selected retail outlets. Up to 85,000 individual sampled quotes are priced each month in approximately 21,000 retail and service establishments. The objective of the Housing Survey is to collect information on cost of shelter and accommodations including type of structure, number of rooms, heating and air conditioning equipment, utility costs, etc. totalling up to 600 pieces of individual information for each sampled housing unit. Approximately 60,000 units are priced in the sample, which is broken into six semi-annual panels.

The price information from both surveys goes through a complex series of editing, estimation and imputation procedures to arrive at price changes (termed price relatives). These price relatives are used to move (inflate or deflate) index values associated with the goods and services in the market basket over time. These indexes are aggregated at various product and geographic levels using additional weighting schemes which represent the relative importance of the commodity within the geographic or product hierarchy which include, at the highest level, food and beverages, housing, apparel, transportation, medical care, entertainment, and miscellaneous goods and services. The highest level of aggregation is the all items index at the national geographic level (typically, this is the CPI reported on the news).

Complexity - What Makes the CPI Difficult

There are a number of intricacies associated with the ongoing production of the CPI. Aside from the statistical complexities of sampling and the shifting patterns of consumer expenditures, there are a number of other problems which necessitate periodic revision, such as shifts in geographic distribution of the population. In addition, there are changes in the market place which require corresponding changes in the sample or the items to be priced. These problems include sample attrition through

retail outlet closure, housing unit demolition, bankruptcies, and product discontinuation, to name a few. These problems are, to some extent, alleviated through sample augmentation and rotation procedures established for both the C&S and Housing Surveys. Other problems include changes in technology and other changes in product quality and quantity. Some of these problems are addressed through the sample actions just described. Others are addressed through item substitution. The implementation of these solutions is accomplished either through procedures instructing field representatives how to address various market situations or through computer based models or both. The sheer volume of information collected requires a large amount of data storage and considerable computer resources to process the data through each stage of processing.

THE 1987 REVISION AND BEYOND

Motivation

As mentioned previously, the CPI mainframe systems were converted to the use of SAS software in the 1987 revision. The seeds of this conversion were sown in early 1981 with a mandate from management within the division responsible for the CPI production system (Trimble, 1982). Motivation for the conversion was partially provided by a lack of vendor support for the existing software language and the concerns expressed by statisticians and economists that the software did not support sufficient numerical accuracy. User access to data was virtually impossible. The system was extremely difficult to modify, and a third or higher generation language was deemed desirable as a replacement. The choice of SAS software from among the candidates was motivated by many considerations. Among these were the ease of learning the language, automatic read/write handling features, improved ease of maintenance, the wide variety of formatting capabilities, the ability to read a wide variety of input data structures, and the availability of preprogrammed PROCs in addition to the DATA step which permitted more traditional procedural processing capabilities. Prototype systems development began at that time with the first major replacement component, the price relative calculation subsystem for the C&S Survey, in place in 1983. The success of this experiment led to the realization that the full conversion was both feasible and desirable.

Structured Systems Engineering and the SAS System

Before discussing the role of SAS software in CPI revision processing, a review of the approach to systems development preceding the implementation of the revision is in order. Generally, it can be described as a traditional life cycle approach as championed by Yourdon and DeMarco in the late 1970s and early 1980s (DeMarco, 1978). Thus, for each system or subsystem that performs processing tasks, there are several phases: requirements definition and analysis, system specification, system design, program design, prototyping,

coding, and testing.

The first stage is the generation of requirements as defined by the users and refined during structured analysis where the software developers improve their understanding of the users' true needs, both stated and hidden. The SAS system is of tremendous help during this process by facilitating the coding of ad hoc programs to examine methodologies and assumptions underlying the users' perceptions of their needs. Once requirements are fully understood, system specification can begin. This process follows structured analysis methodology and results in a functional specification (data flow diagrams and process descriptions) from which the basic data structures at the boundaries of the system can be defined.

From this stage, structured system and program design begins. The logical functional requirements are redefined in terms of system function. Once the overall system design has been determined, program design continues with the objective of a complete physical definition of the processing. This definition in physical terms can result in processes which correspond directly to PROCs or DATA steps, and the communication or movement of data is accomplished via SAS data sets. Prototyping and coding can be accomplished quickly through translation of program design elements into SAS language statements (Mopsik, 1984). This permits illustration of system operation, demonstrates feasibility of design constructs, identifies design weaknesses, and facilitates comparison of alternative coding techniques. The SAS system provides a language which easily supports dynamic data-driven design solutions to processing problems. The use of the SAS Macro facility provides tremendous flexibility in the conditional compilation and execution of code as determined by other results within the program, resulting in more power than can typically be realized in most other traditional programming languages (Henderson, 1982). The Macro facility also provides for an easy modular packaging tool where processing tasks can be isolated into main routines and subroutines which can be controlled using Macro variables as parameters and can be invoked unconditionally or as needed using %IF logic (Phillips, 1985 and 1986).

The testing phase typically follows program design and coding. Several features of the SAS system can facilitate testing. The code used to perform software tests within the program can be controlled by Macro variables. Thus, testing procedures and diagnostics implemented internally in any given program within a system can be activated or deactivated using a single global Macro variable embedded in the production system code. Development of a toolbox of diagnostics, usually written as Macros, can also facilitate testing by elimination of redundant code in favor of generalized code which is parameter driven (Septoff, 1985 and Henderson, 1988). The combination of Macro variables to activate specialized testing paired with a toolbox approach dramatically simplifies the testing process. Hence, as code moves from the development and testing stages

into production, tests established within the design can be reactivated as needed during subsequent system modifications and compared to benchmark results.

The implementation of the 1987 revision required several calendar years and scores of person years in software development activities. For the 1987 revision, the SAS system was effectively used in the software design and implementation process as described above (Johnstone, Rabb, and Sharlin, 1986). This has been beneficial over the years as requirements for system functions change, particularly when the overall design is not significantly affected. To the extent that the SAS software code mirrors the design, the associated DATA and PROC steps corresponding to identified design processes can be lifted and replaced. Modified system interfaces can then be quickly reexamined through the use of PROC CONTENTS to determine if new or modified SAS data set structures correspond to design needs.

System Structure - Platforms, Functions and Software

Having reviewed the general approach to software development, we can turn our attention to the processing cycle for the current CPI production system and the functions performed by applications written using SAS software. In the generation of the CPI on a monthly basis, a number of tasks are performed. This is a cyclical, ongoing process which can be divided into the following six stages: forms generation, data collection, data capture, data editing, data storage, and subsequent imputation/calculation processing to estimate the CPI.

The initial activity, generation of survey forms, is a data-driven process using base SAS software and the SAS Macro facility executed on IBM® 3090™ series computers or equivalents, running under MVS. Forms for both the C&S and Housing Surveys are dynamically generated using a number of data structures which describe items to price (varying by season and location), where to price items (varying by sample rotation schedule) and previous price data. This processing is extremely complex, since each quotation is processed individually, reflecting special pricing conditions, messages to the field representatives, and other unique characteristics. A separate form is produced for each outlet with the items listed in the order in which they are to be priced together with their previous price and related data, using full-page processing and completely blank stock. The forms are printed in the order they are to be used by field personnel. This simplifies work in the field; data collectors merely follow the sequence and record information only when the price has changed from the last data collection. This description is a simplification of the many other actions a field representative may have to take.

Once generated, the forms are mailed to the field offices, where final workload is determined and forms are distributed to data collectors (economic assistants). Data collection is conducted within very specific time periods for both surveys. The forms

are then returned to the national office for data capture and additional processing.

The data capture systems for the C&S and Housing Surveys use a networked collection of 286 and 386 architecture micros with software written in PASCAL and C. The network of micros is linked to a ShareBase 8000 series relational database machine. The data capture systems perform double keyed data entry, associated editing which includes both hard edits for allowable values and soft edits which ensure that the data are consistent across questions on the survey, as well as the resolution of any discrepancies between the two keyed versions. As each screen of data is captured, data are transferred via cable into the appropriate database relations in the ShareBase machine.

Several times a month, the data for a given collection month are uploaded from the minicomputer to the mainframe. This is a complex process requiring communications protocols and a variety of translation functions. The data are eventually stored in a RAPID (from Statistics Canada) database. The use of a RAPID database is well suited for the storage of survey data and provides access to both production users and researchers without the restriction of any given programming language (Turner et al, 1979). In addition, there are BLS-written SAS procedures which provide extract and load capabilities for moving data between RAPID relations and SAS data sets. These procedures include many capabilities to optimize certain recurring query situations.

Review listing software written using the SAS system is executed on the mainframe which permits larger scale consistency checks and review. Expert commodity analysts determine corrective action where data are deemed to be suspect. These corrections are made against the minicomputer database relations and are applied to the mainframe database in subsequent uploads.

Data management features of the SAS system are used widely in a number of CPI systems and subsystems. Above and beyond the interface to the RAPID database, communication within and across subsystems is accomplished via SAS data sets, allowing reliable and easy access and contributing to a uniform system architecture. The intermediate results of subsystems are retained in SAS data libraries permitting faster determination of problem areas and data transport to other subsystems. Additionally, the extensive use of permanently saved SAS data sets provides auxiliary data structures beyond the mainframe and minicomputer databases to assist researchers, predominantly economists and statisticians, in determining improved methodologies and to provide baselines for comparison of results to determine desirability of alternative methods.

Report generation is a large component of the mainframe processing for the CPI production cycle; almost all code for these applications is written using SAS software. There are

hundreds of reports produced for expert review at various stages of processing. The role of commodity analyst review and data correction has already been discussed. The results of each subsystem must also be examined for validity before subsequent activities can begin. To accomplish this objective, a host of report presentations are used varying from the usual PROCs such as PRINT, FREQ, and TABULATE to highly complex custom reports which are data-driven and require sophisticated full-page processing.

Imputation is widespread throughout CPI systems. This type of processing is handled almost exclusively on the mainframe, and a variety of SAS coding solutions have been implemented. Examples of imputation methods employed include allocation, percentage distribution, hot decking, cell collapsing and weighted means. Many subsystems employ a hierarchy of imputation methods where SAS software is employed to examine imputation results and make a determination of the optimal method based on requirements. Perhaps the most complex method is hot deck matching. A generalized implementation approach has been defined which establishes an estimate for a missing value based on other observations with the "closest match" on predefined characteristics using PROC SUMMARY. These and other imputation methodologies are discussed in more detail in previous papers (e.g., Mopsik and Diplo, 1985).

Beyond imputation, millions of calculations and other processing tasks are required to determine individual price movement, relative importance of priced items, and aggregate price movement. These processing tasks include a host of mathematical operations, string operations, logical operations, binary operations, and hexadecimal operations as well as complex recoding/table lookup applications (Johnstone and Ray, 1989) to name a few. Many of these could be accomplished in any traditional programming language, although the built-in features of the SAS system which facilitate such tasks make it a more desirable choice for CPI processing.

Replacement and New Systems - Examples

Since the 1987 revision, there have been surprisingly few replacement systems developed. Generally, changes are implemented into existing systems or new subsystems are written to intercede between existing processing nodes. These changes can be significant, impacting many systems, and are initiated to improve the accuracy and timeliness of the CPI. These changes include the following types of activities: accommodation for sample reduction in both surveys, streamlining the processing cycle for the C&S Survey, reorganization and major changes to the C&S Commodity Review subsystem (used in determination of data corrections), movement of functions from application systems to database functions, elimination of pre-revision transition processing, C&S and Housing Survey collection form changes; an archival system for the massive C&S mainframe database, a sample

sufficiency and augmentation system for the C&S Survey, improved methods for imputation of vacant sampled housing units, and automated methods for noncomparable item substitution in the C&S Survey, to name a few.

One major development effort has been the design and implementation of a Variance Calculation System for the CPI. Following on the heels of an effort to estimate variances of the CPI prior to the 1987 revision (Leaver, Johnstone, and Archer, 1991), this system was developed using a Yourdon/Constantine design approach with strict attention paid to isolation of system functions (Yourdon, 1989). Very highly modular, the system makes heavy use of the Macro facility in packaging work units into modules and explicitly codes intercommunication between modules with Macro variables as parameters. It is not possible to assign a value to a Macro variable in a called module and then return the value as a Macro variable to the calling module via parameter passing. This is not typical of parameter use in other programming languages. The only way to "return" the value is to reference the same Macro variable in both the calling and called module. This was problematic in that it is contrary to the structured design concepts used in the development of the system. The system is highly flexible, relying on user input to determine specifications for estimation such as time period and aggregate level for which estimation is to be performed. It includes many data management capabilities which ensure that duplicate work is not performed if overlapping requests at different times would have resulted in reestimation of values already resident in the database.

In another significant development effort, the Consumer Expenditure Survey installed a major replacement for its publication system in 1991 after several years of development work. The old publication system relied entirely on TPL, the table producing package developed and used at BLS. TPL may be familiar to SAS users as the basis of PROC TABULATE. In the old publication system, TPL was used both to perform a complex series of computations to create the data for the tables and to format the computed values into tabular form. In the new publication system, SAS software is used to compute the values. TPL is used only to format the values into tables. In addition, SAS software is used to generate the needed TPL code automatically, using editable text parameter files as input.

HOW WELL HAS THE SAS SYSTEM FARED

Having described the development and implementation of the production CPI system, we can now address the question of how well development, maintenance and upgrade activities have fared using SAS software. The CPI is an amazingly complex system with many specialized data treatments. Due to changes in methodology and the marketplace, a number of functional changes to processing systems have been required. Many of these changes challenge the original design concepts and affect dozens of subsystems. Managing the process of software redevelopment is done carefully and in a controlled fashion in

the CPI program (Henrickson, 1991).

As we have no comparison language on the same platform engaged in similar processing tasks, we do not have metrics to illustrate and compare relative productivity and resource consumption during development and maintenance activities. Rather, impressions from managers directly responsible for production data processing and the implementation of changes and enhancements were sought to provide an indication of how well the SAS system has been perceived to work in our environment. Generally speaking, the perception is that things have gone quite well given the size and complexity of the system. Indeed, many managers note additional advantages beyond those originally attributed to SAS software in the initial conversion experiments in the early 1980s (described previously). Benefits cited include ease of learning the basics of the language, portability of code to other applications of a similar nature, the self documenting features of SAS data sets, access of data by variable name without knowledge of file layout, automatic read/write operation handling, ability to implement user-written procedures, the wide variety of formatting capabilities, the availability of preprogrammed PROCs, full-page processing, ability to develop toolboxes of multipurpose and reusable code using the Macro facility, and more readable code. As a result, qualified developers become proficient more quickly using SAS software than they do using other languages.

One of the biggest advantages seen on the part of management is the combination of the power of the base language and the Macro facility for systems solutions, together with the accessibility of non-procedural programming (built-in PROCs) for economists, statisticians, and other end users. In addition, managers feel that the SAS system lends itself well to structured programming. This, however, does not alleviate the need for solid design and well maintained documentation. As changing requirements demand the code to stretch beyond the original design constraints, up-to-date documentation becomes critical. To some extent, the readability and seeming understandability of SAS code has encouraged short cuts where documentation is not always maintained on a timely basis. Often, the code itself has had to substitute.

While SAS software is easy to learn and use for routine applications, it can be extremely complicated in the production level systems developed and used in the CPI program. Truly complex operations result in somewhat complex code. This is particularly the case for code written using the Macro facility. While the SAS Macro facility has enabled more powerful and dynamic applications to be written, there are concerns regarding the maintainability of such applications in the face of changing needs. The straightforward nature of the base language disappears when working with the Macro facility. The full workings of applications written using Macro are often obtuse to the novice, and the relationship between the code and the design becomes obscured. This has lengthened the time

required for software redevelopment and testing.

Recruiting and developing staff appropriate for CPI systems development has been difficult. Many of the individuals hired in the systems development offices of the CPI program are recent college graduates. Use of SAS software to develop large-scale production systems or even small scale data processing applications is generally not taught in the computer science curriculum. Expertise is developed through training, but the CPI program has not availed itself of SAS Institute training beyond basic courses. Permanent application development backlogs have also prevented development of a high quality alternative, such as a comprehensive in-house program or a formalized, widespread mentoring program. In the past, the CPI program has experienced significant problems in recruiting individuals who already have SAS software expertise, particularly individuals who also have expertise in building large production level systems.

Software developers within the CPI program have experienced difficulties in the design and testing of systems written using SAS software. This is, in large part, due to a greater reliance on current systems engineering technology and methodologies. This is particularly evident in the lack of software development tools that are applicable to building systems using SAS software. Automated tools such as code generators based on design, complexity analyzers, and testing tools are almost always language specific and do not support code written using SAS software. Even manual structured design and structured testing techniques have to be adapted for use with SAS software. CPI Systems staff have had to develop the methodology, though the proper method of adaptation is not always clear.

Version changes in the SAS system have been time consuming to implement in the CPI production systems. This is primarily due to the highly complex and time critical production systems of the CPI program and the extensive use of BLS-written SAS procedures. Also, version changes to the SAS system have not been completely upward compatible, requiring thorough inspection, modification, and testing of production programs. This has required and continues to require a considerable investment in time and personnel. In spite of this cost, the benefits of SAS version changes, particularly the new features and capabilities provided by SAS Version 6 have made conversion efforts worthwhile. Examples include SELECT/WHERE processing, PROC SQL, and data set compression and indexing.

In summary, the use of SAS software has not presented overwhelming impediments. The problems in finding resources with the level of expertise required combined with a lack of software development tools to assist designers has resulted in a perception that system development using SAS software is not as easily accomplished as was thought several years ago. Still, the benefits are seen as far outweighing these difficulties, and

the choice of the SAS system for software architecture is still regarded as fundamentally sound.

VISIONS FOR THE FUTURE

Other Uses of SAS Software - Research and Analysis

In both research for a better understanding of the market place and for improved computational algorithms and estimation methods used in the CPI, SAS software plays a significant role. While the production system and software redevelopment efforts are still wedded almost exclusively to the mainframe platform, researchers are beginning to make much more significant use of SAS software on microcomputers. Currently, SAS release 6.04 for DOS is available on the BLS network, and there are over 200 active users of the basic product throughout the Bureau. Most of the statistical researchers addressing issues related to the CPI use SAS software on 386 machines. Typically, the mainframe is being used for extracting data from the RAPID databases and performing basic data shaping using SAS software. The Micro-to-Host Link is then used to facilitate a download to the individual's machine. This is accomplished, in part, through a RabbitGate SNA/SDLC gateway that links nodes on the network to the mainframe via high speed data transmission lines. Once the data are resident on the researcher's machine, investigative analysis and modeling activities are performed to address research questions at hand. These vary from sampling research and applications to investigation of estimation techniques. The advantages of the SAS system cited elsewhere in this paper are generally seen to hold, although there is still preference to use a variety of languages and packages depending on the individual researcher's preferences and background. Another interesting development is the construction of initial models and related code on the PC and the subsequent porting of the application to the mainframe to lower timesharing costs. The availability of Version 6 of SAS software on the mainframe has improved the possibility of portable applications. Still to be addressed within the Bureau is the development and availability of user friendly procedures to permit multiplatform processing and intercommunication, sparing the user from generating the required script using SAS/CONNECT™ software. At this point, individual offices must provide sufficient expertise to accomplish these tasks.

In a prototype application, SAS/AF® software is being considered for use to provide an interactive analysis system for commodity analysts. Termed the Professional Review Analysis subsystem, it is intended to provide the capability to access both the mainframe database (currently through RAPID user interface) and minicomputer database (using IKON, an in-house developed retrieval system that builds SQL against ShareBase databases) and amalgamate data into SAS data sets. The system then provides the analyst a host of ways to present the data in graphic, tabular and other comparative approaches. The intent is to improve the ability to analyze and understand changes in

the market place and samples without the analyst having to do any programming. Given that IKON fully interfaces with Microsoft's Windows operating system, returning queried data to a given window, potential recipients of the prototype system are looking forward to the Windows release of SAS 6.07 in hopes of improved point and click analysis capabilities.

Ongoing Development - SAS/AF Software Use

The CPI program currently uses a Production Control System (PCS) to control submission of batch jobs on the mainframe. The PCS was developed within BLS using the command language of the mainframe editor available at the commercial data center used by the CPI program. The PCS prompts the user for a series of job submission parameters and does symbolic substitution (similar to that of the SAS Macro facility) in a JCL template. The PCS then allows the user to submit the job with the desired parameters already in place in the JCL.

The PCS has served the CPI program well, but it does have some problems. Not every data center has the mainframe editor command language on which the PCS is based. A number of problems have occurred when the CPI program has required data centers to install the mainframe editor, including system performance problems and lack of support from data center personnel because the product is unfamiliar to them. In addition, the PCS is a very limited interface; basically it is an old style, line-by-line command based editor that can not support many desirable user interface techniques. For example, to change a parameter after it has been entered, the user has to ask explicitly to make a change. The user is then led through the change process by yet another series of prompts.

To address these problems, the CPI program is currently developing a new production control system, PCS/AF, based on SAS/AF software. In the new PCS/AF system, each application will have associated with it a customized screen displaying a skeleton of all of the necessary job submission parameters. The user will be able to use the cursor to move around the screen, entering and changing parameter entries at will before electing to submit the job. The capabilities of the old PCS will be retained, but the user interface will be greatly improved. Each new application will require only the development of a job parameter entry screen. PCS/AF will be extensible either through modification of the core program or through modification of individual screens. Use of SAS/AF software will not require data centers to install additional software, since any data center that wishes to provide automated data processing services to the Bureau must already offer SAS software and support. In addition, PCS/AF applications may be ported to micros to minimize mainframe activities. Finally, PCS/AF will provide a user friendly interface that will reduce operator fatigue and errors. The CPI program has already developed a working PCS/AF prototype for one of the program's major subsystems. Work is underway to extend PCS/AF for use in the rest of the CPI program's subsystems

(Phillips and Peer, 1992).

Conversion to a Commercial RDBMS and Use of New Interfaces

As mentioned previously, the CPI program uses the RAPID database management system in almost all of its mainframe subsystems. RAPID is a sophisticated product offering basic data management functions, a wealth of useful utilities, and excellent capabilities for storing statistical and survey data. As described previously, the interface between the SAS system and RAPID is provided by a series of BLS-written SAS procedures that were developed in the CPI program. RAPID is not a commercially available product, and the CPI program's copy, as well as the procedures that provide the interface with the SAS system, are maintained solely in house. Maintenance in house has allowed the CPI program to achieve superior levels of optimization and has also allowed the creation of new data types specially suited to the CPI program's data. Maintenance and support of RAPID and the procedures that provide the interface with the SAS system are, however, highly complex and have sometimes proven very difficult. Moreover, RAPID is an old system that is not truly relational and that does not support a Structured Query Language (SQL) consistent with current standards.

To address these problems, the CPI program intends to replace the current database manager with a commercial relational database management system (RDBMS). Using a commercial RDBMS will eliminate entirely the need to maintain the system and support software. Using a commercial RDBMS would also reduce the need for user-written interface products. The SAS system has built-in interfaces for many major RDBMS products. The availability of a SAS system interface will greatly facilitate the conversion process; source code changes may involve little more than replacing the calls to BLS-written SAS interface procedures with DATA steps or calls to built in SAS procedures through products such as SAS/ACCESS® software. In addition, the availability of PROC SQL should dramatically increase the capabilities of CPI production processing systems. Improved database access within the scope of the SAS system will inevitably enhance end user computing such as research and analysis activities. The CPI program is currently engaged in a pilot project examining conversion of mainframe database structures to DB2®, the Bureau standard for a mainframe RDBMS, in order to determine the scope and magnitude of such a project.

Proposed 1997 CPI Revision Systems - Examples

Since planning activities for the next revision are at an early stage, it still remains unclear what the full scope of the 1997 revision will be. The objectives of the survey and the mission of the program as a whole remain unchanged, as does the overall objective of improved accuracy and timeliness. The cycle of information to be gathered remains fundamentally the

same, though the methods of collection, capture, and subsequent processing of data are likely to undergo significant changes. Some of these proposed changes in the form of potential systems are discussed below.

The collection of data in the field is anticipated to be computer assisted. Termed Computer Assisted Data Collection or Computer Assisted Survey Information Collection, the process is envisioned to be accomplished via notebook or pen-based computing. While this topic is beyond the scope of this paper, it is worthy of note in that SAS software will likely not play a role other than, perhaps, in workload scheduling. Furthermore, a significant proportion of systems development staff time may be spent on survey instrument software development and a host of data communication issues to transport data from the field through intermediary minicomputers or directly to national office minicomputers for assessment and correction and thence to the mainframe databases. Thus, development of other processing systems may be less ambitious in their scope.

Data capture being performed at the point of collection by personal computer devices will eliminate the need for forms based systems. Thus, the time and resources consumed in forms generation, data capture and preliminary data editing are viewed as diminishing in the new environment. The implementation of an alternative RDBMS has been previously discussed, though there will still be the necessity of moving data from minicomputer to the mainframe database, and new upload procedures may be required.

The mainframe portions of the CPI processing system, the Housing Survey in particular, are seen as undergoing additional major renovation above and beyond the potential conversion of the mainframe RDBMS used. Some likely Housing Survey systems to be written and implemented using SAS software are described below.

First, a sample must be selected. While the sample selection itself may follow roughly the same model as used in the previous revision, the identification of primary and all shadow (alternate) samples will result in a data structure to be later used by a Sample Maintenance and Control System (SMCS). This SMCS, to be developed using SAS software, is envisioned to include not only a master RDBMS database with all sample related information from sample selection and initiation, but subsidiary pricing databases which contain survey information for those samples being actively priced. As a sample approaches deficiency, the SMCS will alert appropriate reviewers who will then have a variety of sample actions available. The SMCS, under statistical review and verification, will have the capability to optimally select augmentation or rotation samples on a nonscheduled basis automatically. This is seen as improving the statistical validity of the CPI in a more time responsive manner. The SAS system is also seen as playing a significant role in reformulation of major estimation systems, such as the price relative calculation system. In this

and other computation systems, significant changes in estimation models, methods and algorithms are projected, primarily for the Housing Survey.

Thus, while plans remain unfixed at this time, it is apparent that the SAS system will remain the bedrock of mainframe software architecture in the 1997 CPI with many new or replacement applications systems to be developed. Above and beyond systems mentioned and others yet to come, the possibility of introducing multiplatform processing using SAS software will be explored.

Potential Role of Distributed Processing

Several factors suggest the possibility of a more distributed processing environment for the production system for the CPI. Among these are the projected increasing costs of mainframe processing, the proliferation of 386 and higher architecture machines available, the procurement of high speed workstations to perform other processing, the portability of Version 6 SAS application software across operating systems and platforms, and improvements in interplatform communication. Given these factors, the possibility of isolating certain subsystems to selected platforms and the intercommunication of parallel processing results introduced by subsystem independence are seen as viable solutions to lowering overall costs of processing as well as improving the timeliness of the CPI. There are a number of issues to be considered in such an endeavor, many of which have not been addressed as yet. Should this distributed processing become a reality, it is likely that SAS software will play an even more prominent role in the CPI production system with its increasing power, flexibility, and capabilities realized through MultiVendor Architecture™ and expanding products.

CONCLUSIONS

Originally chosen as an alternative to traditional programming languages in selected applications, SAS software became the cornerstone of mainframe software architecture for the production processing systems of the CPI. Despite more recent methodological problems in the implementation of structured design techniques using SAS software and the relatively high cost of version changes, the decision is seen as remaining fundamentally sound given the benefits derived from the choice. It is anticipated that the CPI program will continue to use SAS software in the 1997 revision and beyond and that the program will be increasingly able to take advantage of new products and features in a multiplatform environment as they become available.

ACKNOWLEDGEMENTS

The authors wish to acknowledge all of the employees of the Bureau of Labor Statistics and contractors who provided assistance in the development of this paper. They include, but are not limited to, the following: Stephen G. Wright, Division

Chief of Consumer Prices and Consumption Studies (CPCS) who quietly insisted this paper get done; Karen Hassmer, Edward Pratt, Donald Jones, and Richard Schroeder, CPCS managers who gave of their time to discuss their impressions and experiences; Paul Armknecht, Assistant Commissioner for Consumer Prices (CPI program office) who ensured other groups were aware of our efforts; Ken Stewart and David Barry who met with us to discuss the professional review analysis subsystem prototype; the late Curtis Jacobs, former Division Chief of Statistical Methods Division within the CPI program office; Brian Hedges, Acting Division Chief of the Statistical Methods Division within the CPI program office, for facilitating discussions with his staff; Janet Williams, Branch Chief of CPI research, and her staff members, Dr. Sylvia Leaver, David Swanson, Eugene Brown, Bob Baskin and Shawn Jacobson who discussed their research projects with us. Our special thanks to all who reviewed various drafts and submitted valuable comments and suggestions.

AUTHOR CONTACT

The authors may be contacted at:

Bureau of Labor Statistics
Office of Technology and Survey Processing
Directorate of Survey Processing
Division of Consumer Prices and Consumption Studies
Room 7600
Bicentennial Building
600 E. St. N.W.
Washington, D.C. 20212

DISCLAIMER

The views presented herein are solely those of the authors. Impressions and conclusions reached do not necessarily reflect agency thinking or beliefs. In addition, future matters are the conjecture of the authors, and while they may represent current thinking on the part of some individuals in the agency, they do not reflect a commitment or definitive direction on the part of the CPI program, its offices or personnel.

REFERENCES

- Bureau of Labor Statistics, BLS Handbook of Methods, Volume II, 1984, U.S. Government Printing Office, Washington, D.C.
- Bureau of Labor Statistics, BLS Handbook of Methods, Volume II, 1988, U.S. Government Printing Office, Washington, D.C.
- DeMarco, Tom. Structured Analysis and System Specification, 1978, Yourdon, Inc., New York, N.Y.
- Dippo, Cathryn S. and Jacobs, Curtis A., "Area Sample Redesign for the Consumer Price Index," American Statistical

Association, 1983 Proceedings of the Survey Research Methods Section, Washington, D.C.

Henderson, Don. "SAS Tutorial: Conditional Execution of DATA and PROC Steps," SUGI '82 Proceedings, SAS Institute, Cary, N.C.

Henderson, Don. "A Toolbox Approach to Networking SAS Experts or Re-using instead of Re-inventing the Wheel", SUGI '88 Proceedings, SAS Institute, Cary, N.C.

Henrickson, Elizabeth. "The Afterlife of a Production System: Software Library Maintenance in a Large Data Processing Environment," SUGI '91 Proceedings, SAS Institute, Cary, N.C.

Johnstone, James and Rabb, Merry and Sharlin, Joshua. "The Use of the SAS System in the Development of the Production System for the Consumer Price Index (CPI)," SUGI '86 Proceedings, SAS Institute, Cary, N.C.

Johnstone, James and Ray, Craig. "A Survey of Table Lookup Techniques for the SAS System," SUGI '89 Proceedings, SAS Institute, Cary, N.C.

Leaver, Sylvia and Johnstone, James and Archer, Ken. "Estimating Unconditional Variances for the U.S. Consumer Price Index for 1978-1986", American Statistical Association, 1991 Proceedings of the Survey Research Methods Section, Washington, D.C.

Mopsik, Judith. "The SAS System Can Replace Traditional Programming Languages," SUGI '84 Proceedings, SAS Institute, Cary, N.C.

Mopsik, Judith and Dipbo, Cathy. "The Data Adjustment for the Consumer Expenditure Survey," American Statistical Association, 1985 Proceedings of the Statistical Computing Section, Washington, D.C.

Phillips, Jeff. "Designing Macro Based Systems," SUGI '85 Proceedings, SAS Institute, Cary, N.C.

Phillips, Jeff. "The Use of the Macro Facility in a Systems Development Environment," SUGI '86 Proceedings, SAS Institute, Cary, N.C.

Phillips, Jeff and Peer, Todd. "Are We Having Fun Yet? Human Factors in Designing a SAS/AF Job Submission Application," SUGI '92 Proceedings, SAS Institute, Cary, N.C.

Septoff, David. "Developing and Using Macro Utility Libraries," SUGI '85 Proceedings, SAS Institute, Cary, N.C.

Trimble, Harvey. "Can SAS Replace Traditional Programming Languages in Large Systems Development? A Manager's Perspective," SUGI '82 Proceedings, SAS Institute, Cary, N.C.

Turner, M.J., Hammond, R., and Cotton, P., "A DBMS for Large

Statistical Databases," 1979 Proceedings of the 42nd Biennial Congress of the International Statistics Institute, Manila, Philippines.

Yourdon, Edward. Modem Structured Analysis, 1989, Yourdon Press, Eaglewood Cliffs, New Jersey.

SAS, SAS/ACCESS, SAS/AF, MultiVendor Architecture, and SAS/CONNECT are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. DB2, IBM, OS/2, and 3090 are registered trademarks or trademarks of International Business Machines Corporation.

Other brand and product names are registered trademarks or trademarks of their respective companies.