VPLX

Program Documentation

Volume 1: Introduction

Robert E. Fay
U.S. Bureau of the Census
February 25, 1998

Table of Contents

Appendices:

Preface

History. This documentation describes VPLX, a general program for variance estimation from complex samples. VPLX was begun in 1989. The initial and continuing objective is to develop a general purpose program for the analysis of survey data generally and Census Bureau products specifically. The program has been used for a number of purposes at the Census Bureau. It is the production system for the variance calculations for the Current Population Survey (CPS), the source of the monthly unemployment rate and other labor force statistics. A number of other researchers have also used the program. It is freely available from the Census Bureau's web site.

This draft marks a renewed effort to document the program. A substantial predecessor draft, which continues to be available, has remained incomplete since 1995. The challenge of explaining the program and experience with users' questions convinced the author that fundamental simplifications in the syntax were required to make the program more accessible. Hence, the author's priorities shifted from completing the 1995 documentation to documenting an improved system.

From the beginning, the PC has been the primary development environment. The PC is well suited to editing and testing. It is also expected to remain one of the primary platforms for VPLX applications, particularly by users outside the Census Bureau. During some periods in VPLX's past history, however, further development has also been confined by the PC environment. Until 1997, the current version remained compatible with Windows 3.1 but at the cost of restricting the complexity of the code. In particular, concerns for resources limited previous effort to improve the program. Windows 95/NT appears to have effectively removed these constraints. The Fortran code, written within the FORTRAN 77 standard, continues to port to other systems, in both UNIX and VMS, at the Census Bureau.

Knowledge of Fortran is not specifically required to use VPLX. Instead, users prepare sets of instructions to VPLX in the *VPLX language*. The syntax of the VPLX language is the principal focus of a substantial revision of the program, begun in earnest in April 1997. Both the original VPLX syntax and the revision share some common elements of other familiar programming languages but do not follow one specific model. The revision of the VPLX language has incorporated specific features of Fortran and the SAS language more closely than the original syntax, however. It is not as powerful a language as Fortran, C, SAS, or other general languages, but it does support a number of complex calculations familiar in the analysis of sample survey data.

This new documentation will attempt to keep pace with the revision of the syntax. It describes the new syntax where it has been implemented but continues to employ the old syntax for the remaining parts. Specifically, the most basic element of the system, the CREATE step, has now been revised, as has the REWEIGHT step. This documentation shows the new CREATE step syntax used in conjunction with the older DISPLAY step syntax. The TRANSFORM step is next slated for major revision, and new documentation of this step will follow. Subsequently, the DISPLAY step will be revised and expanded, and DISPLAY step applications appearing in the documentation will then be updated accordingly. Consequently, with the exception of the DISPLAY step applications and a restricted use of the current TRANSFORM step to illustrate the new REWEIGHT syntax, this new documentation describes the revised system. Applications written following the new CREATE and REWEIGHT syntax will not become obsolete as a result of any of the planned revisions to other steps of the program.

Although this documentation will evolve to reflect the new syntax, the program will continue to support legacy applications written in the previous syntax. A VPLX application may mix steps where some steps are written in the old syntax and some in the new, as long as the two forms are not mixed within a step.

Although knowledge of Fortran is not required to learn VPLX, prior experience with some computer language provides a critical foundation. This documentation is generally not an adequate first introduction to concepts such as reading and writing files, sequential processing of cases, double precision arithmetic, representation of quantities by variable names, *etc*. At the Census Bureau, the majority of VPLX users lack Fortran experience but have used SAS, and this background is quite helpful for learning VPLX. The author believes that experience with SPSS, S, or other statistical systems that use a programming language would also provide a sufficient background.

For the last several years, releases have been dated according to *yy.mm* where *yy* is the last 2 digits of the year and *mm* denotes the month. In recognition of the approaching millennium, the form has now shifted to *yyyy.mm.* Considerable effort has been made to make the changes upwardly compatible so that previous VPLX applications run identically on new versions. Conversion to the new syntax has required a few incompatibilities with versions prior to 97.09, and Appendix A1 lists these.

Organization of the Documentation. The new documentation is divided into three volumes:

C *Introduction* (This volume) Except for this preface, the emphasis is on describing core features of the VPLX system. This volume is written to be read or skimmed in order, since generally chapters build upon and presuppose material in previous chapters. More complex or less important features of VPLX are occasionally mentioned but accompanied by a reference to sections in the other two volumes rather than described in complete detail.

C *Replication Methods* This second volume discusses selection of replication methods and file organization for implementation in VPLX. The first chapter is a relatively nontechnical introduction, but subsequent chapters provide a detailed description of requirements and considerations in selecting a specific replication method for an application.

C *Advanced Features* This volume is organized by topic. A first chapter summarizes the scope of each of the subsequent chapters. Except where otherwise explicitly noted, each chapter is written to be read independently of the others. In general, however, the volume presupposes familiarity with the *Introduction*.

This organization reflects an effort to serve three types of users, recognizing that a given individual may wear different hats on different occasions.

C *Analysts* These users typically bring questions to data. When the files and replication methods have been chosen in advance by a statistician, *Introduction to VPLX* should provide all the necessary detail for most applications.

C *Statisticians* These users have the responsibility of selecting a replication method for a survey. Generally, they are expected to have some familiarity with the first volume but to use *Replication Methods* as a guide to set up a replication application for analysts. In some cases, they or collaborators may employ REWEIGHT and other techniques described in *Advanced Features* to set up replicate weights.

C *Advanced Users* Some applications will require features of VPLX beyond the capabilities described in *Introduction to VPLX.* Such users are likely to require topics included in *Advanced Features.* This volume is designed to be read as the need arises rather than systematically.

Although the details appear in the appendices, readers are encouraged to install VPLX and the examples before attempting to study the documentation in detail. In many cases, the documentation has extracted key points from an example, but there is more to be learned from the complete version. Readers are invited to experiment with any of the examples.

It is also hoped that many readers will find that they can successfully skim sections rather than study them in full detail. The annotated examples illustrate most of the important points.

Computing Environments. The development of VPLX has been strongly influenced by the available resources in the Census Bureau's computing environment. The three primary systems with which there is the largest accumulated experience are:

- IBM-compatible PC's, generally at or above the 486 level, with 16 MB (preferably 32 MB) of RAM, running Windows 95 or NT. From 1989 until mid-1993, development work was primarily on 286-level PC's running under DOS. From then until 1997, the target machines were at or above the 386 level with 8 MB running under Windows 3.1. The Windows 3.1 version remains available but does not support the new features.

  The new Windows 95/NT version of VPLX is available as a single .EXE from the Census Bureau's web site. Appendix A2 provides instructions on installing VPLX.

- The VAX VMS environment. The Census Bureau acquired several DIGITAL machines under a procurement initiative for the 1990 Census, and these machines continue to be used in developing the 2000 census. Other important applications such as the processing for the monthly labor force characteristics from the CPS have used this environment. A number of key economic surveys and censuses also rely on this environment.

- UNIX-based workstations. Specifically, VPLX has been installed and used on some of the Census Bureau's SUN and Digital workstations. These workstations have been used for both research and production for demographic surveys, including the CPS, and the Survey of Income and Program Participation (SIPP). The Fortran source of the UNIX version of VPLX is available from the Census Bureau web site. Appendix A3 provides guidance on installing the program under UNIX.

There is essentially no experience on IBM mainframes at this date. The Census Bureau does not have such machines available for development and testing. An important obstacle is the development of a means for users to access files under the more complex specification requirements of MVS, TSO, *etc.* The current version of VPLX includes some code to handle an extended syntax to address these issues, but the work has not yet been finished. The author's current assumption is that, except for the issue of file specification, the portable FORTRAN 77 code should be almost transparently portable to the IBM environment or others supporting this standard

The Fortran source code is available to users who want to attempt compilation and testing on other environments. For example, no attempt has yet been made to compile and test the program on MACINTOSH systems, but the system has not been developed with any intention to exclude any environment with an adequate FORTRAN 77 compiler and hardware resources.

Why VPLX? VPLX is not alone as a general system for the computation of variances from complex samples: SUDAAN of the Research Triangle Institute, WesVarPC of WESTAT, and PC-CARP developed under the leadership of Prof. Wayne Fuller of Iowa State University, are among the alternatives. Alan Zaslavsky provides a general review of available systems at

http://www.fas.harvard.edu/~stats/survey-soft/survey-soft.html.
Nonetheless, there are important distinctions among these as to approach, portability, and capabilities. The author's belief is that, at the current time, the availability of several such systems should help to raise the standard of practice in the analysis of complex survey data.

VPLX does appear to occupy an important niche: as a replication-based system with the potential to estimate the variance of highly complex estimators such as those exemplified by many of the Census Bureau's most important statistical products, including the decennial census, the CPS, and others. The emphasis on complex applications provides the motivation for a language-based system rather than one relying on a graphical user interface.

Although comparisons among systems are useful, the author does not expect a single system to emerge as the standard. Large institutions are likely to benefit from supporting their own system, which can more readily adopt to their needs, rather than relying on an external source. For example, Statistics Canada has a variance system, the Generalized Estimation System, closely tied to their estimators and other parts of their processing systems.

A second question should also be answered - should the Senior Mathematical Statistician of the Census Bureau continue to devote time to a computer program? Again, the author's opinion is in the affirmative. VPLX has become a primary tool for the author for inventing and testing new statistical methodology.

As I understand, Virginia has one of the lowest fees for "vanity" license plates and has the highest per vehicle participation. My car carries:

VPLXER

I.P.6

Acknowledgements. Although I have benefitted from the experience of working with a number of users, four have provided especially useful advice: Carol Gunlicks, Sylvia Leaver, Chris Moriarity, and George Train. Others have supported and encouraged this effort, including Daniel Kaspryzk, William Butz, Robert Groves, Robert Tortora, and Cynthia Clark.