

# 1. Introduction

## 1.1 General Description

VPLX is a FORTRAN program for the estimation of variances from complex sample surveys through replication methods. Replication methods, in general, obtain an estimate of sampling variance by carrying out the estimation for the full sample on a series of subsamples of the data, and the program is specifically designed for this purpose.

Most Census Bureau and other government sample surveys, as well as many other national surveys, employ complex sample designs. One consequence is the typical use of weighted data in estimation. Many widely available statistical systems, such as SAS, SPSS, and others, are able to use the survey weights provided on public use files to reproduce the official estimates. Typically, however, the variance estimates provided by these systems do not consider the effect of the complex design on the reliability of the estimates. In some situations, variance estimates based on an assumption of simple random sampling may be roughly corrected by rules of thumb. For many purposes, including the Census Bureau's own research, however, more accurate variance estimates are required. For this purpose, the sample design generally must be considered. VPLX and some other specialized systems, such as SUDAAN, now provide a means to do so for a variety of designs and estimation problems.

VPLX has also proven valuable to carry out some other kinds of calculations associated with the analysis of survey data. This consequence was not originally foreseen in the initial development of the program. For example, VPLX can be used, in some cases, to carry out complex survey weighting procedures as a check of other systems. Its language for the manipulation of multi-dimensional arrays may make VPLX a possible choice for some applications. VPLX is also sometimes useful for Monte Carlo evaluations of statistical estimators. Currently, VPLX insists on carrying out at least one replicate sample in addition to the full sample, requiring users to work around this restriction. Intended future development of the program will remove this awkwardness, improving VPLX's utility in these other areas.

Knowledge of FORTRAN is not required to use VPLX. Instead, users prepare sets of instructions to VPLX in the *VPLX language*. The primary purpose of this documentation is to describe this language. The VPLX language shares some common elements of other familiar programming languages but does not follow any one specific model. It is not as powerful a language as FORTRAN, C, SAS, or other general languages. Nonetheless, it does support a number of calculations familiar in the analysis of sample survey data.

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. Appendix B is provided specifically to assist SAS users in understanding differences between the VPLX and SAS language. The author believes that experience with SPSS, S, or other statistical systems would also provide a sufficient background. Persons whose principal computer experience is limited to word processing and spreadsheets are unfortunately likely to find learning VPLX challenging.

VPLX was begun in 1989 and has been freely available to users on one or more Census Bureau computer systems since 1991. The program continues to grow. For the last few years, releases have been dated according to *yy.mm* where *yy* is the last 2 digits of the year and *mm* denotes the month. Considerable effort has been made to make the changes upwardly compatible so that previous VPLX applications run identically on new versions. Bugs have been corrected, however, and generally a new version has been issued to correct any bugs affecting the computed estimates or standard errors. The subject of errors is discussed further in Section 1.5.

## 1.2 VPLX Files

A central feature of the VPLX system is the *VPLX file*. This term stands for a complex file structure employed by VPLX to organize the data required for variance estimation through replication. The VPLX file holds both data and *metadata*, characteristics of the data such as choice of replication method, variable names, labels, and other attributes.

A VPLX application is organized into a series of steps. Parallels between VPLX and the SAS system are possible here: for example, a user may assemble a SAS data set from raw input data with the data step and then run one or more of the available SAS procs. SAS steps may be run relatively independently of each other because the relevant information is communicated between steps as metadata included with the SAS data set. Similarly, VPLX steps are almost completely independent and communicate with each other through the contents of VPLX files.

The design of the VPLX file has undergone far fewer changes than the program, although the addition of new features may force a redesign of the VPLX file within the next few years. Again, an attempt will be made to maintain upward compatibility in implementing this revision. As of February, 1995, the current file design is 92.03, but the program still accepts 90.04 as input to most steps.

In addition to VPLX files, VPLX applications often require reading and writing from several different files. Two files are particularly important: the character file containing the VPLX language commands prepared by the user, which will generally be called the *input command file* in the documentation; and an output file, called the *print file*, to which VPLX generally echoes commands and reports results. In practice, of course, users may usually prefer to inspect the print file in an editor rather than physically printing it.

### 1.3 Computing Environments

The development of VPLX has been strongly influenced by the available resources in the Census Bureau's computing environment. The 3 primary systems with which there is the largest accumulated experience are:

- IBM-compatible PC's, generally at or above the 386 level, with 8 MB or more of memory. Until mid-1993, development work was primarily on 286-level PC's running under DOS. By 1993, however, the DOS 640K environment was beginning significantly to constrain further development.

Conversion to the Microsoft FORTRAN Powerstation compiler overcame this obstacle. The compiler runs under Windows but produces 32-bit executables that run under DOS with a DOS extender. The extender allows VPLX to take advantage of extended memory. Current development and testing primarily occurs on 486-based machines and the resulting source code is periodically ported to other environments. This version requires 8 MB of memory.

Almost all non-Census Bureau users of VPLX use the PC version, which is sent as an executable file that does not require compilation, along with files associated with the DOS extender.

- The VAX VMS environment. The Census Bureau acquired several DIGITAL machines under a procurement initiative for the 1990 Census, and important applications such as the processing for the monthly labor force characteristics from the Current Population Survey (CPS) have moved to this environment. A number of key Economic Surveys and Censuses also rely on this environment.
- UNIX-based workstations. Specifically, VPLX has been installed on some of the Census Bureau's SUN workstations and tested on others, including an HP and a Digital workstation based on the Alpha chip.

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 can be made 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.

#### **1.4 About This Documentation**

The present version of the documentation represents a substantial reorganization and revision since a series of chapters were drafted or completed around May, 1993. VPLX includes several features. Years of working on applications alone and with other users, and experience teaching the program to classes at the Census Bureau, led to the conclusion that some VPLX features are far more essential than others. This version, although still attempting to cover useful features, is specifically designed to group topics by importance. Chapters 3 and 4, for example, are virtually required reading for almost any application. Almost all advanced applications depend on Chapter 7. Each chapter begins with a section attempting to indicate the relative importance of the chapter. To the extent that users may read only half or less of this documentation and still be able successfully to carry out their analysis, this reorganization will have succeeded.

Most chapters include endnotes. In general, endnotes, when they appear in the text, offer further explanation on specific points but generally can be skipped by most readers with little loss. If the reader is puzzled by a statement in the main text at the very point at which an endnote is indicated, the reader should feel invited to consult the additional information.

#### **1.5 Checking VPLX Results for Accuracy**

A previous program written by the author, CPLX, was released in 1988 and appears never to have required correction for bugs. Compared to VPLX, however, CPLX was highly limited in scope. The FORTRAN source for version 94.11 includes almost 39,000 lines, more than 10 times those of CPLX. Unfortunately, some bugs have been encountered in previous VPLX releases.

As new features are added, they are initially tested, but the testing may not identify a problem that only occurs under specialized conditions. At any one point, VPLX includes many undocumented features that the author is continuing to test alone or by working directly with other users. During this period, the syntax or generality may be considerably revised without notice. For example, `BINARYREAD` and `BINARYWRITE`, still undocumented, are likely targets of further development. Once a feature is documented or develops a user community, however, changes are generally restricted to upwardly compatible enhancements.

By the time a feature is released for widespread use, relatively few errors are subsequently encountered. Nonetheless, computer users should be alert to the possibility that a specific result from VPLX could be in error.

All general computer languages invite the possibility of error. Common programming mistakes in implementing the VPLX language are a more likely source of error in VPLX applications than error in the VPLX program itself. All users, and new users in particular, are encouraged to check estimates from VPLX against results obtained from other systems, particularly when there is an opportunity to do so. As a general rule, once the estimates have been checked, then the variance estimates from VPLX are virtually always accurately computed if the information on the sample design and other information required by the replication method is accurately provided. For example, VPLX estimates should be checked against SAS estimates, when they are available, but users should generally not expect the variance estimates from VPLX to match those from SAS.

Chapter 11 further discusses strategies for debugging complex VPLX applications.

## **1.6 Why VPLX?**

VPLX is not alone as a general system for the computation of variances from complex samples: SUDAAN of the Research Triangle Institute, WESVAR of WESTAT, PC-CARP developed under the leadership of Prof. Wayne Fuller of Iowa State University, TREES by Prof. David Bellhouse are among the alternatives. 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 systems is the most likely route 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 Current Population Survey, and others.

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 is developing a variance system closely tied to their estimators, even though they are not as far along on this work as some other 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 fundamental tool for the author for inventing and testing new statistical methodology.

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



**VPLXER**