Texas A&M University Census Research Data Center

Computing Environment and Available Software

Overview

Researchers using restricted access data sets must conduct their analyses within the computing environment of the CRDC facility. These notes briefly describe the nature of that computing environment and the software that is available for performing analyses.

Computing Environment

CRDC facilities provide Unix-style workstations that are networked with secure servers at various locations (e.g., the Center for Economic Studies at the Census Bureau in Suitland, Maryland). The workstations run the Red Hat edition of Linux, a variant of the Unix operating system. The programs available in the CRDC environment are versions specific for the Linux operating system. In some cases, there may be minor differences in how these programs are used in comparison with usage of versions for the PC and Mac operating systems.

Restricted access data do not reside within the CRDC facility itself. They reside on Census and other computer systems elsewhere and are accessed from the CRDC workstation via a secure network connection.

Available Software

The CRDC environment provides a range of general purpose and specialized software for performing data management and statistical analysis. The following is a brief list of the programs available in CRDC computing environment.

General statistical packages	SAS, Stata, and R
Special purpose statistical packages	HLM, SUDAAN
Specialized languages	MatLab, Gauss
GIS-Spatial analysis software	GRASS, GeoDa

Short descriptions of these programs are provided below along with links to websites where additional information is available.

General Statistical Packages

SAS. SAS is a statistical package which can handle large datasets and features analysis of variance, mixed models, regression, categorical analysis, Bayesian analysis, multivariate analysis, survival analysis, psychometric analysis, cluster analysis, nonparametric analysis, survey data analysis, and multiple imputation for missing values.

http://www.sas.com/technologies/analytics/statistics/stat/

STATA. Stata is a general-purpose statistical software package created in 1985 by StataCorp. It is used by many businesses and academic institutions around the world. Most of its users work in research, especially in the fields of economics, sociology, political science, and epidemiology.

http://www.stata.com/

R. R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R. R provides a wide variety of statistical (linear and nonlinear modellng, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

http://www.r-project.org/

Specialized Statistical Analysis Packages

HLM. Hierarchical linear and nonlinear models (also called multilevel models) have been developed to allow for the study of relationships at any level in a single analysis, while not ignoring the variability associated with each level of the hierarchy. The HLM program can fit models to outcome variables that generate a linear model with explanatory variables that account for variations at each level, utilizing variables specified at each level. HLM not only estimates model coefficients at each level, but it also predicts the random effects associated with each sampling unit at every level.

http://www.ssicentral.com/hlm/index.html

SUDAAN. SUDAAN® is an internationally recognized statistical software package that specializes in providing efficient and accurate analysis of data from complex studies. SUDAAN is ideal for the proper analysis of data from surveys and experimental studies, since SUDAAN procedures properly account for complex design features, such as correlated observations, clustering, weighting, and stratification.

http://www.rti.org/sudaan/page.cfm/About_SUDAAN

Specialized Languages for Analysis and Computing

MATLAB. MATLAB® is a high-level technical computing language and interactive environment for algorithm development, data visualization, data analysis, and numeric computation. You can use MATLAB in a wide range of applications, including signal and image processing, communications, control design, test and measurement, financial modeling and analysis, and computational biology.

http://www.mathworks.com/products/matlab/

GAUSS. The GAUSS Mathematical and Statistical System is a fast matrix programming language widely used by scientists, engineers, statisticians, biometricians, econometricians, and financial analysts. Designed for computationally intensive tasks, the GAUSS system is ideally suited for the researcher who does not have the time required to develop programs in C or

FORTRAN but finds that most statistical or mathematical "packages" are not flexible or powerful enough to perform complicated analysis or to work on large problems.

http://www.aptech.com/gauss.html

GIS-Spatial Analysis Software

GRASS. The Geographic Resources Analysis Support System, commonly referred to as GRASS, is a free Geographic Information System (GIS) used for geospatial data management and analysis, image processing, graphics/maps production, spatial modeling, and visualization. GRASS is currently used in academic and commercial settings around the world, as well as by many governmental agencies and environmental consulting companies. GRASS is an official project of the Open Source Geospatial Foundation.

http://grass.itc.it/

GEODA. GeoDa is the latest incarnation of a collection of software tools designed to implement techniques for exploratory spatial data analysis (ESDA) on lattice data.1 It is intended to provide a user friendly and graphical interface to methods of descriptive spatial data analysis, such as autocorrelation statistics and indicators of spatial outliers. The design of GeoDa consists of an interactive environment that combines maps with statistical graphics, using the technology of dynamically linked windows.

http://geodacenter.asu.edu/