

StatSoft®

Business White Paper

***STATISTICA Data Miner:
Integrating R Programs
into the Data Miner Environment***

Last Updated: June 2003

U.S. Headquarters: StatSoft, Inc. • 2300 E. 14th St. • Tulsa, OK 74104 • USA • (918) 749-1119 • Fax: (918) 749-2217 • info@statsoft.com • www.statsoft.com

Australia: StatSoft Pacific Pty Ltd.
Brazil: StatSoft Brazil Ltda.
Czech Republic: StatSoft Czech Rep. s.r.o.
France: StatSoft France

Germany: StatSoft GmbH
Hungary: StatSoft Hungary Ltd.
Israel: StatSoft Israel Ltd.
Italy: StatSoft Italia srl

Japan: StatSoft Japan Inc.
Korea: StatSoft Korea
Netherlands: StatSoft Benelux BV
Poland: StatSoft Polska Sp. z o. o.

Portugal: StatSoft Iberica Ltda.
Russia: StatSoft Russia
Singapore: StatSoft Singapore
S. Africa: StatSoft S. Africa (Pty) Ltd.

Spain: StatSoft Espana
Sweden: StatSoft Scandinavia AB
Taiwan: StatSoft Taiwan
UK: StatSoft Ltd.

Table of Contents

Executive Summary	3
Introduction to <i>STATISTICA Data Miner</i>	3
Introduction to R	5
Introduction to <i>STATISTICA Data Miner</i> Customization	5
Example: Integrating an R Script as a Custom Data Miner Node	7
Setting Up Your Environment.....	7
Integrating an R Script as a Custom Data Miner Node	7
Conclusions	16

Executive Summary

StatSoft, provider of the **STATISTICA** product suite, is committed to partnering with our customers to provide comprehensive analytics solutions to fit their business needs. **STATISTICA Data Miner** provides a comprehensive suite of data mining, analysis and visualization tools all within a single software platform.

STATISTICA Data Miner provides an open architecture for utilizing third party components and/or custom applications developed internally within your organization. **STATISTICA Data Miner** provides an integrated development environment and over 11,000 functions within an Application Programming Interface (API), exposed as Visual Basic for Applications (VBA) from within **STATISTICA**.

This paper provides an overview and an example for utilizing **STATISTICA** analyses and graphics in tandem with scripts developed with the R programming language. The R language is popular in academia and in bioinformatics. R analyses and procedures can be utilized within **STATISTICA Data Miner** and incorporated into the **Data Miner** environment as “custom nodes.” This approach is useful in a number of scenarios, one of which is in the utilization of the **STATISTICA Data Miner** platform for the analysis of genomics data, where there are many bioinformatics algorithms written in R (for additional details, see <http://www.bioconductor.org>).

Introduction to **STATISTICA Data Miner**

STATISTICA provides the most comprehensive and effective system of user-friendly tools for the entire data mining process - from querying databases to generating final reports.

- **STATISTICA** contains a **comprehensive selection of data mining methods**;
- A selection of comprehensive, **complete data mining projects (solutions), ready to run, and set up to competitively evaluate alternative models** (using bagging (voting, averaging), boosting, stacking, meta-learning, etc.), and to produce presentation-quality summary reports;
- An extremely easy to use, **drag-and-drop based user interface** that can be used even by novices, but is still highly flexible, customizable, and provides one-click access to the underlying scripts;

- **Powerful, interactive data exploration (drilling, slicing, dicing) tools**, including the most comprehensive selection of **interactive, exploratory graphics-visualization tools**;
- Ability to handle/process multiple data streams simultaneously;
- **Optimized for processing extremely large data sets** (including options to pre-screen even over a million variables, and/or draw stratified or simple random samples of records using DIEHARD-certified random sampling procedures);
- **Highly optimized read (and write) access to large databases**, including IDP (In-Place Database Processing) technology that reads data asynchronously directly from remote database servers (using distributed processing if supported by the server), and bypassing the need to “import” data and create a local copy;
- **Flexible deployment engine**, integrated with a custom development environment allowing you to manage optimized analytic objects (nodes) for data mining using quick, industry standard, Visual Basic scripts (VB is built into the system);
- **Extremely fast and efficient deployment** via portable, XML syntax based PMML (Predictive Models Markup Language) files for prediction, predictive classification, or predictive clustering of large data files; trained models can be shared between desktop and Client/Server installations;
- **Options to write predicted values, classifications, classification probabilities, prediction residuals**, and so on directly into external databases for subsequent analyses, selection, etc.; by using efficient IDP (In-Place Database Processing) technology for reading and writing information from and to external databases, datasets of extremely large sizes can be analyzed and scored (i.e., used to update predicted values, classification probabilities, and so on in the database);
- **Open, COM-based architecture, unlimited automation options, and support for custom extensions** (using industry standard VB (built in), Java, or C/C++/C#);
- **Multithreading and distributed processing architecture** delivers unmatched performance (offered in the Client-Server version) including **super-computer-like parallel processing technology** that optionally scales to multiple server computers that can work in parallel to rapidly process computationally intensive projects.

- **Complete Web-enablement** (via **WebSTATISTICA** offering support for all data mining operations, including the interactive model building, via Internet browser using any computer connected to the Web); this ultimate enterprise data analysis/mining system allows you to manage projects over the Web and work collaboratively "across the hall or across continents."

Introduction to R

Immediately below is an excerpt from the R Web site (<http://www.r-project.org/>) that provides an introduction to the scope and purpose of the R language:

"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."

The R Foundation (<http://www.r-project.org/>) is a not for profit organization setup to develop and support the R platform. R is a freely available software under the terms of the Free Software Foundation's GNU General Public License (please see <http://www.r-project.org/COPYING>).

- StatSoft does not develop, re-sell or support R. StatSoft does recognize that there are **STATISTICA** users who also use R and have invested in the development of procedures and analyses built on the R language.

Introduction to **STATISTICA Data Miner** Customization

The industry standard **STATISTICA** Visual Basic language (integrated into **STATISTICA**) offers incomparably more than just a "supplementary application programming language" that can be used to write custom extensions. **STATISTICA** Visual Basic (SVB) takes full advantage of the object model architecture of **STATISTICA** and allows you to access programmatically every aspect and virtually every detail of the functionality of the program. Even the most complex analyses and graphs can be recorded into Visual Basic (SVB) macro programs and later be run repeatedly or edited and used as building blocks of other applications. **STATISTICA**

Visual Basic adds an arsenal of more than 11,000 new functions to the standard comprehensive syntax of Microsoft Visual Basic thus comprising one of the largest and richest development environments available.

The **STATISTICA Data Miner** is an open architecture system: All nodes (except for the input spreadsheet node, which is used during interactive operations to connect to a **STATISTICA** data file or connections to external databases for in-place processing) are **STATISTICA** Visual Basic scripts that access the vast functionality of **STATISTICA** analytical and graphics routines. However, this interface also permits users to develop their own analytic nodes, either using entirely **STATISTICA** Visual Basic scripts (following the industry standard Visual Basic conventions), or only using **STATISTICA** Visual Basic to "connect" their own external subroutines and procedures written in any computer language, or provided by other vendors.

The general architecture of the program is very simple, but extremely flexible and powerful: Data are connected (or generated) via data acquisition nodes, which define an object called the *InputDescriptor*. This object contains all information about the input data source, the nature of the selected variables (e.g., categorical dependent variables, continuous predictors, etc.), codes (e.g., to identify the groups or classes for categorical or class variables), and other information (e.g., information about censoring, and so on). The *InputDescriptor* (or multiple *InputDescriptors*) is (are) the data sources for and can be connected to subsequent analyses and computations, which themselves can create *InputDescriptors* for further analyses. For example, a multiple regression node may take as the input an *InputDescriptor*, perform the analyses, create various output documents such as graphs, etc., and also produce a data set with predicted values as the *InputDescriptor* for further analyses; hence, subsequent analyses can be connected to this node to perform further analyses with the *InputDescriptor* computed by the multiple regression node.

There are different types of nodes to perform different tasks (e.g., analyses, subsampling and filtering, etc.). Each node is fully defined by two files: A **STATISTICA** Visual Basic script (.svx file) and a data miner initialization file (.dmi) that contains information about the functional properties of the respective node, descriptions of the computations that are performed, default parameter values for the respective analyses, etc.

This paper assumes a general familiarity with the approach to **STATISTICA Data Miner** customization. For additional details, please consult the electronic documentation provided with **STATISTICA**, specifically the topic entitled "How to Write .svx Scripts for Data Miner."

Example: Integrating an R Script as a Custom Data Miner Node

This example illustrates how to integrate an existing R script within **STATISTICA Data Miner** and make it available as a custom node that can be used within all Data Miner projects. In this simple example, we will utilize R to create two or more arrays filled with random numbers from a normal distribution. We will get summary statistics from R and create histograms within **STATISTICA** using these data.

Setting Up Your Environment

Before beginning to replicate this example, you must have R installed on either the same computer running **STATISTICA Data Miner** or a computer accessible from the **STATISTICA** server. If you do not have R installed, please visit the Comprehensive R Archive Network (CRAN) currently at <http://cran.r-project.org/mirrors.html>. This example demonstrates the integration of **STATISTICA** via the COM interface to R. You will also need to download and install the COM server for R, currently available at <http://cran.r-project.org/contrib/extra/dcom/>. Documentation is included with both the R distribution and the R (D)COM server. Please refer to it for installation instructions, platform and compatibility notes, and the latest troubleshooting information.

Integrating an R Script as a Custom Data Miner Node

The integration of an R program involves the creation of SVX (scripts) and DMI (description) files. Within **STATISTICA**, open the macro development editor available from Tools > Macro > **STATISTICA** Visual Basic Editor. From the Tools menu, select References. Confirm that the R D(COM) ("StatConnector") libraries are included in the project.

The **STATISTICA** Visual Basic program utilizes the StatConnector object provided within the R D(COM) interface. The complete code for the example SVX file is provided below:

```
'=====
==
' 1. Short program description for node browser
'=====
==
' Random Variables Generated by R
' Input parameters:
'     NumberOfRandomVariables : Number of random variables
'     NumberOfRandomCases    : Number of random cases
' Results nuggets:
'     Summary statistics spreadsheet
'     Histograms
```

```

' Output data: Spreadsheet of the random numbers generated by R
' Deployment: None

'=====
==
' 2. Complete program description and comments.
'=====
=
' Summary of Random Variables Generated by R
'
' Copyright: StatSoft, Inc.
' Date: June. 12, 2003
'
' Input: DataIn() As InputDescriptor
' Output: DataOut() As InputDescriptor
'
' The node is a SVB (STATISTICA Visual Basic) subroutine defined by
the
' system. It has two parameters which specify DataIn and DataOut
' respectively. The DataIn() is an InputDescriptor object array
' specifying the InputSpreadsheet, variable selections, and some other
' parameters. DataOut() is an InputDescriptor object array which is
' generated by the node and could be used as DataIn() in following
analysis.
'
' This node acquires random variables generated by R and outputs the
' summary statistics and histograms.

'#Uses "*CommonDataMinerInputErrorMessages.svx"
Option Explicit
Option Base 1

Private Sub AnalysisNode( _
    DataIn() As InputDescriptor, _
    ByVal ReportDocs As StaDocCollection, _
    DataOut() As InputDescriptor)

    Dim NumberOfRandomVariables As Long
    NumberOfRandomVariables=Dictionary("NumberOfRandomVariables")
    Dim NumberOfRandomCases As Long
    NumberOfRandomCases=Dictionary("NumberOfRandomCases")
    Dim GraphType As Integer
    GraphType = Dictionary("GraphType") '1 Regular, else Multiple

    Dim x As StatConnector
    Dim rnorm As Variant, stats As Variant
    Dim ss As Spreadsheet, sum As Spreadsheet
    Dim g() As Graph
    Dim s As String
    Dim sym As String, sname As String
    Dim r As Double
    Dim I As Integer, J As Integer
    Dim Lower As Integer, Upper As Integer
    Dim evaString As String

```

```

On Error GoTo handle_error
Set x = New StatConnector

x.Init ("R")

evaString = "r <- matrix(rnorm(" & Str(NumberOfRandomCases) & "*" &
Str(NumberOfRandomVariables)
      &"), ncol = " & Str(NumberOfRandomVariables) & " )"
x.EvaluateNoReturn (evaString)
rnorm = x.GetSymbol ("r")
Lower = LBound(rnorm)
Upper = UBound(rnorm)

Set ss = New Spreadsheet
ss.SetSize(Upper-Lower+1,NumberOfRandomVariables)
ss.Header = "Random numbers generated using R"
For I = Lower To Upper
  For J = 0 To NumberOfRandomVariables-1
    ss.Value(I+1-Lower,J+1) = rnorm(I,J)
  Next J
Next I
For J = 1 To NumberOfRandomVariables
  ss.VariableName(J) = "Random number " + Str(J)
Next J
ss.AutoFitVariables

x.EvaluateNoReturn ("stats <- summary(r)")
stats = x.GetSymbol ("stats")
Lower = LBound(stats)
Upper = UBound(stats)

Set sum = New Spreadsheet
sum.SetSize(Upper-Lower+1,NumberOfRandomVariables)
sum.Header = "Summary of random number variables generated using R"
For J = 1 To NumberOfRandomVariables
  sum.VariableName(J) = "Random number " + Str(J)
Next J
For I = Lower To Upper
  For J = 1 To NumberOfRandomVariables
    sum.Value(I+1-Lower,J) = I+1-Lower
    sum.SetTextLabel(J,I+1-Lower,stats(I,J-1))
  Next J
Next I
sum.AutoFitVariables
sum.AutoFitCase

DrawHistogram ss, g, GraphType

ReportDocs.Add sum
If GraphType = 1 Then
  For J = 1 To NumberOfRandomVariables
    ReportDocs.Add g(J)
  Next J

```

```

Else
  ReportDocs.Add g(1)
End If

ReDim DataOut(1 To 1)
Set DataOut(1) = DataIn(1).Clone(ss,scClone)

no_error:
  x.Close
  Exit Sub

handle_error:
  If x Is Nothing Then
    MsgBox ("R Object not available")
  Else
    x.Close
    MsgBox ("R evaluation error")
  End If

End Sub

Private Sub DrawHistogram (ss As Spreadsheet, g() As Graph, gType As
Integer)
  Dim newanalysis As Analysis
  Dim vars() As Long
  Dim I As Long, nvars As Long

  On Error GoTo StaError

  nvars = ss.NumberOfVariables
  ReDim vars( 1 To nvars )
  If gType = 1 Then
    ReDim g( 1 To nvars )
  Else
    ReDim g( 1 To 1 )
  End If
  For I = 1 To nvars
    vars(I) = I
  Next I
  Set newanalysis = Analysis (sc2dHistograms, ss)
  With newanalysis.Dialog
    .Variables = vars
    If gType = 1 Then
      .GraphType = scHistogramRegularPlot
    Else
      .GraphType = scHistogramMultiplePlot
    End If
    .FitType = scHistoFitNormal
    .ShowingType = scStandard
    .BreakBetweenColumns = False
    .ShowPercentages = False
    .YAxisOption = scLeftNumber
    .DisplayDescriptiveStatistics = False
    .DisplayKolmogorovSmirnovTest = False
  End With

```

```

        .DisplayShapiroWilkTest = False
        .DisplayTotalCount = False
    End With

    If gType = 1 Then
        For I = 1 To nvars
            Set g(I) = newanalysis.Dialog.Graphs.Item(I)
        Next I
    Else
        Set g(1) = newanalysis.Dialog.Graphs.Item(1)
    End If

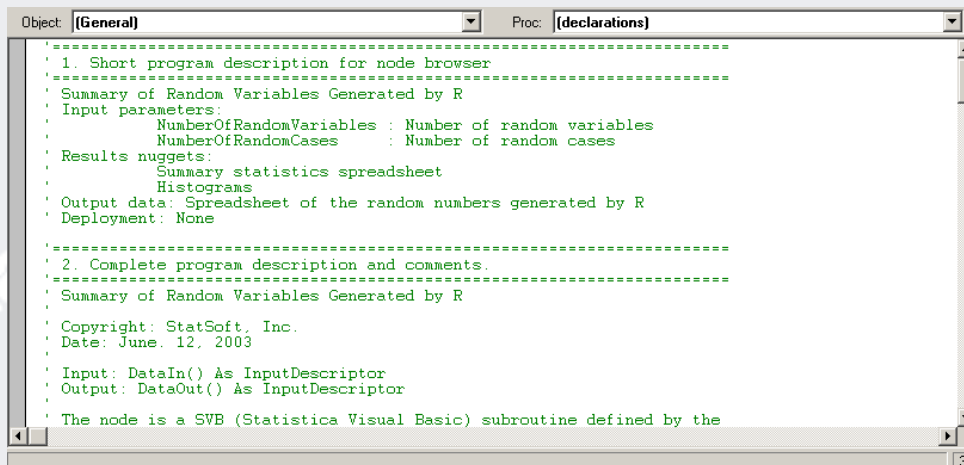
    On Error GoTo 0
Exit Sub

StaError:
    MsgBox "STATISTICA running error."
Exit Sub

End Sub

```

Copy and paste this script into the SVB Macro Editor as shown below:



Save the script as RandomVariablesByR.svx.

It is necessary to create a Data Miner descriptor file (DMI) for deployment within **STATISTICA Data Miner** and use across all potential Data Miner projects. The content of the DMI file for this example is provided below:

```

# Summary of Random Variables Generated by R
[NODE]
TYPE=2
NAME="Random Variables Generated by R"
GROUP="Input Data and Data Acquisition"
ICONANALYSISID=scUnknown
MACRO="RandomVariablesByR.svx"

```

```

LONGDESC="This node acquires random variables generated by R and
outputs the summary statistics and histogram graphs."
SHORTDESC="Summary of Random Variables Generated by R"
WEBIMAGE=SubsetIcon.gif

[PARAM1]
NEWTAB="General"
NAME=NumberOfRandomVariables
TYPE=Integer
MIN=1
MAX=100
INITIALVALUE=2
LONGNAME="Number of random variables"
DESCRIPTION="Specifies the number of random variables to generate by
R."
USERACCESS=READWRITE

[PARAM2]
NAME=NumberOfRandomCases
TYPE=Integer
MIN=1
MAX=10000
INITIALVALUE=1000
LONGNAME="Number of random cases"
DESCRIPTION="Specifies the number of random values in each variable
generated by R."
USERACCESS=READWRITE

[PARAM3]
NAME=GraphType
TYPE=Integer
MIN=1
MAX=2
INITIALVALUE=1
ENUM="1 Regular histogram | 2 Multiple histogram"
LONGNAME="Type of histogram"
DESCRIPTION="Specifies the type of histogram generated by STATISTICA."
USERACCESS=READWRITE

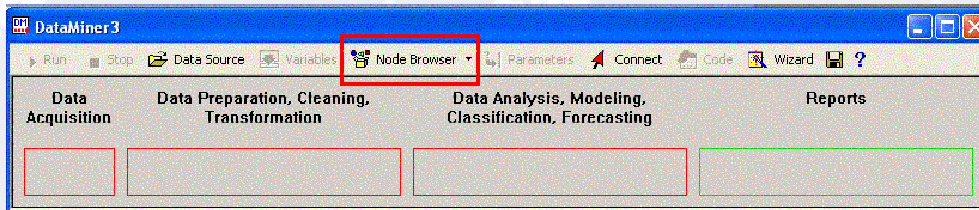
```

Open a text editor, paste in the above content, and save the file as RandomVariablesByR.dmi.

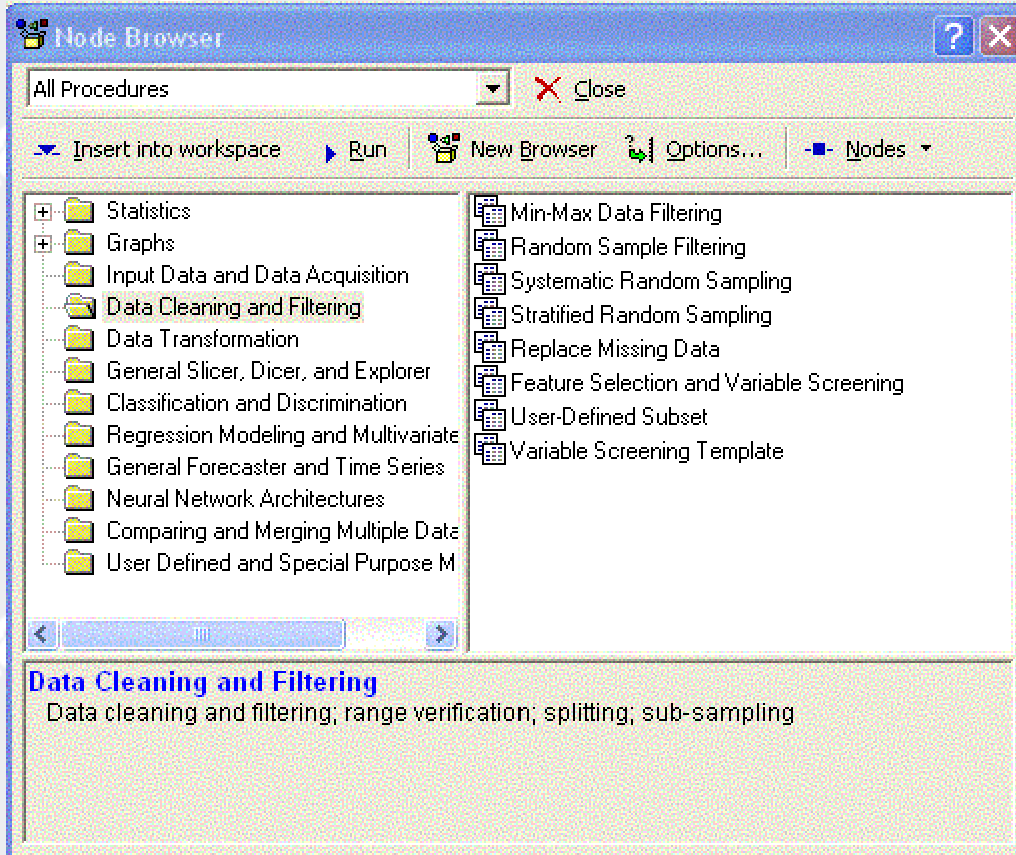
Copy the two (2) files to the following location on the server or workstation running **STATISTICA Data Miner**:

```
[Drive Letter]:\[STATISTICA DATA MINER HOME]\Data Miner\
```

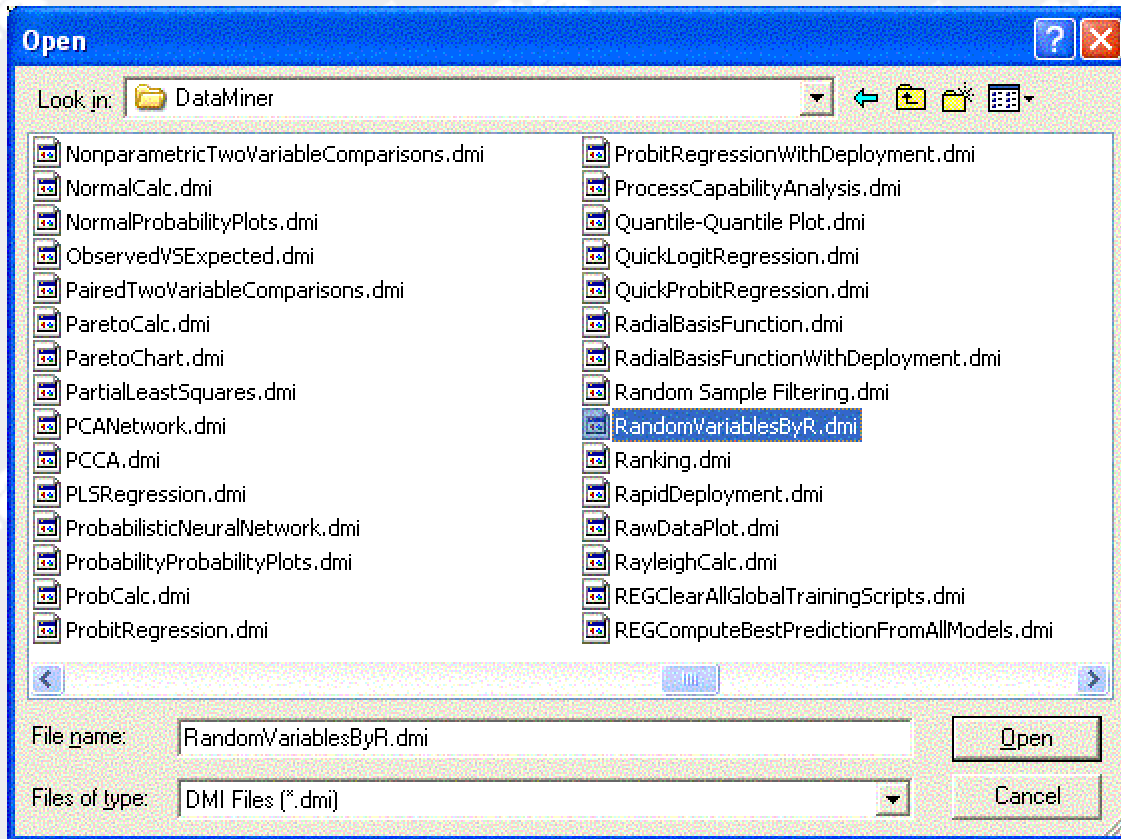
Next, within **STATISTICA Data Miner**, select *Statistics > Data Miner > All Procedures* from the pull-down menus.



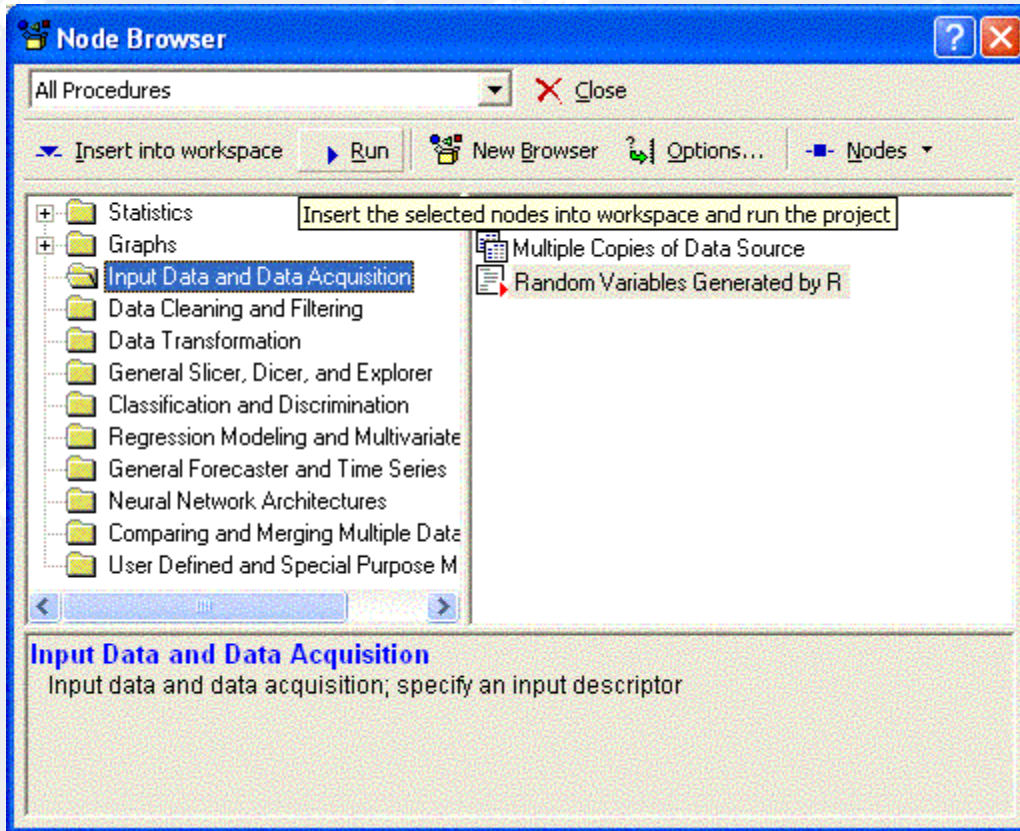
To import the new node, click on the *Node Browser* button within the *Data Miner* project user interface.



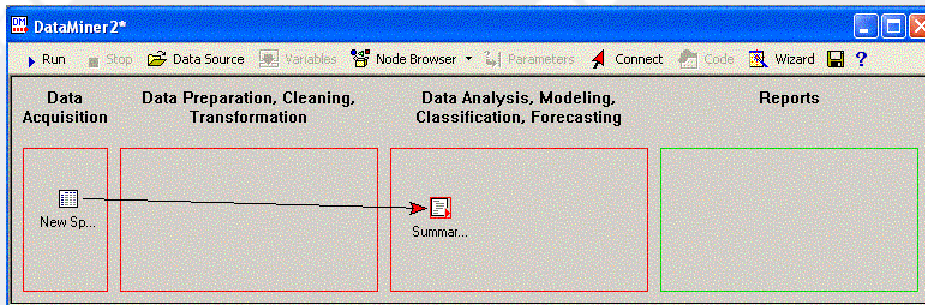
Select *Nodes > Import New Node*. Browse to the folder with the DMI file and select it.



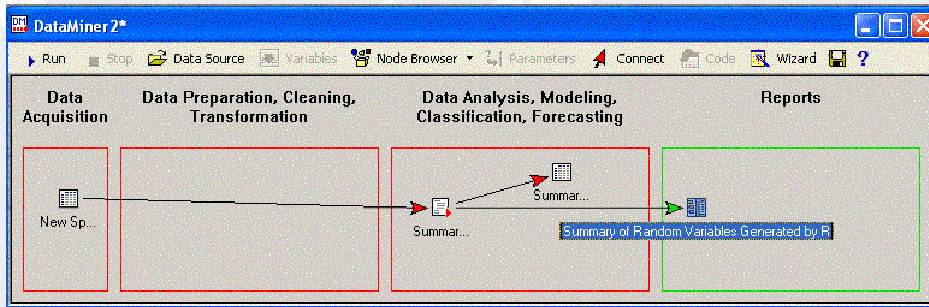
When you return to the *Nodes* dialog, the new node is available within the *Input Data and Data Acquisition* category.



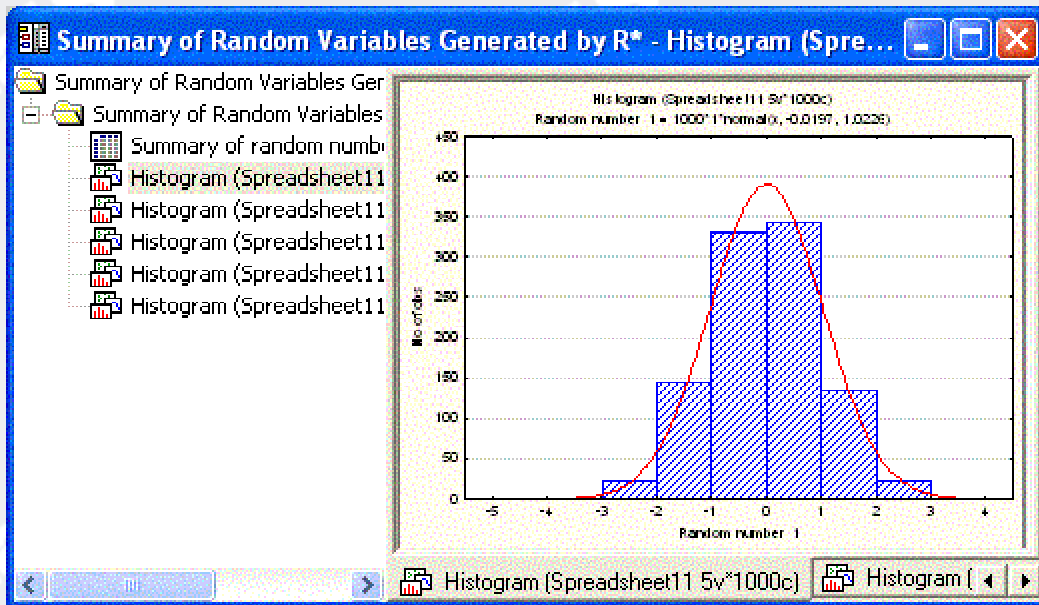
Double-click on the node labeled *Summary of Random Variables Generated by R* and close the *Node Browser*. The node will be included in the *Data Miner* project.



Open a Data Source (via the *Data Source* button) and connect it to this new node. Press *F5* to run the project. When the node is run, it will interface with the R (D)COM server, create a StatConnector object, pass to it the R script(s), and then run *STATISTICA* procedures on the results.



Double-clicking on the *Reports* node will display a **STATISTICA** Workbook with descriptive statistics and histograms for the results.



Conclusions

STATISTICA Data Miner provides a powerful, flexible and easy-to-use software platform for the data mining, analytics and visualization needs of your organization. This paper provided an introduction to the features and benefits of **STATISTICA** along with an example of its open architecture from which third party applications and scripts may be integrated with the wealth of **STATISTICA** procedures.

For more details about **STATISTICA Data Miner** and for a customized demonstration of the system for your needs, please contact StatSoft at 918-749-1119.