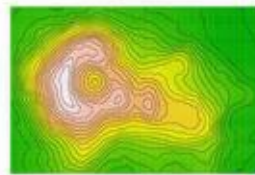
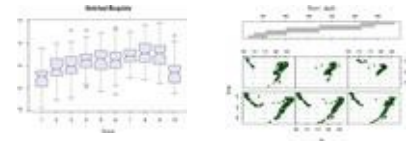
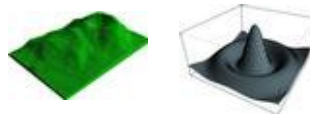
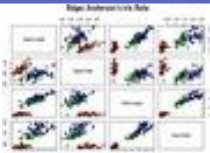


Hans W. Borchers
DECRC/I4

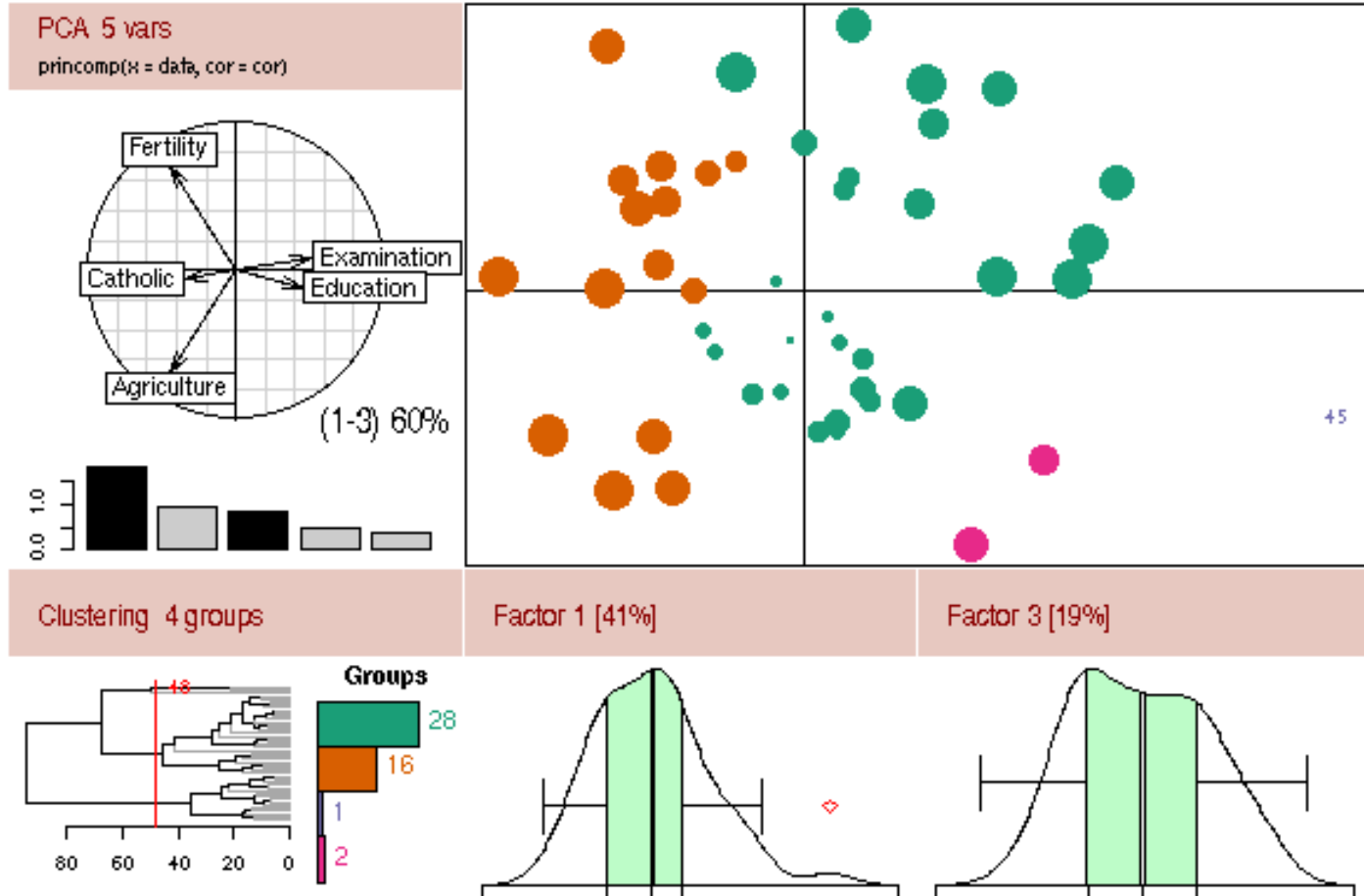


The *R* Statistics Package

A Short Introduction



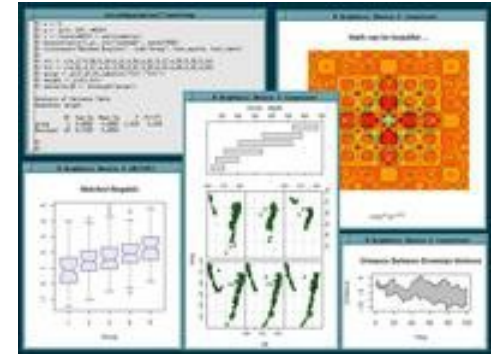
Example R Desktop



What is *R* ?



- *R* is a **language** and **environment** for statistical computing and graphics.
- *R* is available as **Free Software** under the terms of the GNU General Public License (GPL, but not LGPL).
- *R* Foundation of Statistical Computing
- *R* is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes
 - an effective data handling and storage facility,
 - a suite of operators for calculations on arrays, in particular matrices,
 - a large, coherent, integrated collection of intermediate tools for data analysis,
 - graphical facilities for data analysis and display either on-screen or on hardcopy, and
 - a well-developed, simple and effective programming language.



R is 'GNU S'



- **R** comprises an implementation of the **S** language (for statistical programming):

ACM Award 1998

- developed 1976-80 at **AT&T Bell Labs** by **J. M. Chambers** and co-workers
- quasi-**standardized** since 1991/92 (as "New S")
- implemented in **S-PLUS** (commercial) and **R** (free)
- runs on all Windows, all Mac, all Linux versions
- modestly **object-oriented** -- comparable to Matlab (development of a much stronger OO approach under way)
- excellent documentation through **on-line help** (help files in HTML, LaTeX, PDF, Windows Help, Unix man pages)
- graphical user interfaces through **Tcl/Tk** integration
- easily extendible through **C** and **Fortran** routines (DLLs under Windows)
- (D)**COM** interface under Windows (Excel, VB, .NET to come)
- S Repository at CMU (StatLib)

The Brown Book
The Blue Book
The White Book
The Green Book



R Packages



- About 20 base packages, 1000s of functions

base boot class cluster ctest eda foreign grid KernSmooth lattice lqs MASS methods
mgcv mle modreg mva nlme nls nnet rpart spatial splines stepfun survival tcltk
tools ts

- More than **250** add-in packages

abind acepack adapt **ade4** agce akima amap AnalyzeFMRI anm ape ash aws Bhat bindata blighty
bootstrap bqtl brlr car **cclust** cfa CGIwithR chron CircStats clim.pact cmprsk cobs CoCoAn coda
combinat conf.design cramer date Davies DBI dblcens **deal** deldir Devore5 diamonds dichromat
dispmod dr dse **e1071** effects ellipse emplik EMV evd exactRankTests fastICA fdim fields fracdiff
g.data gafit gbm gee geepack GeneSOM genetics GenKern geoR geoRglm gld GLMMGibbs
glmmML gpplib grasper GRASS gregmisc gss gstat gtkDevice haplo.score hdf5 hier.part homals
hwde ifs ineq ipred ismev ISwR KMSurv **knnTree** lars lasso2 **leaps** lgtdl lntest locfit logspline
lokern lpridge **maptree** Matrix maxstat mclust1998 **mclust** MCMCpack mda meanscore mimR mix
mlbench moc MPV msm muhaz multcomp multidim multiv mvnmlc mvtnorm ncomplete netCDF
NISTnls nlrq norm normalp normix noverlap npmc Oarray odesolve oz pamr panel pastecs pcurve
pear permax pinktoe pixmap pls.pcr **polspline** polynom princurve pspline PTak qtl quadprog
quantreg qvcalc R2HTML RadioSonde RandomFields **randomForest** RArclInfo Rcmdr
RColorBrewer relimp rgenoud rimage rmeta RmSQL **RMySQL** RODBC ROracle rpvm RQuantLib
rsprng RSQLite RsvgDevice Rwave sampfling SASmixed **scatterplot3d** segment sem serialize
session sgeostat sm sma sn snow sound SparseM spatstat spdep splancs StatDataML statmod
strucchange subselect SuppDists survey survrec systemfit tensor tkrplot tree tripack **tseries**
twostage varddiag vcd vegan VLMC VR waveslim **wavethresh** wle **xgobi** XML xtable



Web Pages around R



- The [R Project](#) for Statistical Computing and [CRAN](#), the Comprehensive R Archive Network
 - ◊ The R-Project [R-Help](#) and [R-Packages](#) mailing lists web interfaces
 - ◊ The [R News](#) and [ASA's Statistical Computing & Graphics](#) Newsletter
 - ◊ The [Omega](#), [Bioconductor](#), and [Rmetrics](#) Projects for Statistical Computing
 - ◊ [ESS](#), *Emacs Speaks Statistics* (cf. [ESS and XEmacs for Windows Users of R](#))
 - ◊ [R and S-Plus Packages, Functions, and Documentation](#) with the Hmisc and Design packages and the S-Plus/R Function Finder
 - ◊ Ripley & Venables' [MASS3](#) page with [R contributions](#), [Rwin](#), and [Rtools](#)
 - ◊ [S Poetry](#), a free book on advanced S programming, and [Rtips](#)
 - ◊ [RCC](#), the R to C compiler at Rice University
 - ◊ [R Web Interfaces](#)
 - ◊ [Interactive Software](#) developd at University of Augsburg (Klimt, Mondrian, JGR)
 - ◊ [GGobi](#) Software for R, and the [Statlib S Archive](#) at CMU, Pittsburgh
 - ◊ [Graphical Models](#) and software related to the gR project, especially the [mimR](#) web site, the [BUGS Project](#) (WinBUGS), and the [openBUGS](#) website
 - ◊ [RPy](#) (R from Python) and [RSPython](#) (R/SPlus - Python Interface) | [R for Ruby](#) Library and [Tools/R](#) Ruby bindings for R (Approximity)
 - ◊ [SVMlight](#), [lp_solve](#) and yet another [yacas](#) computer algebra system
 - ◊ Statistical quality control package [QToolBox](#)
 - ◊ [RWinEdt](#) and the [SWinRegistry](#) package
 - ◊ [R GUIs](#), see also [SciViews-R](#) with [Tinn-R](#) editor
 - ◊ Software for [Extreme Value Theory](#) in R and S
 - ◊ Software by [Duncan Murdoch](#) (and others)
 - ◊ [Jim Lindsey's R page](#)

Data Mining in R



<u>Method</u>	<u>Function</u>	<u>Package</u>			
NN	nnet	<i>nnet</i>	SVM	svm	<i>e1071</i>
lvq	lvq, olvq	<i>class</i>	KNN	knn, knn1	<i>class</i>
SOM	SOM	<i>class</i>		knntree	<i>knntree</i>
CART	rpart	<i>rpart</i>	PCA	pca	<i>mva</i>
	tree	<i>tree</i>	discriminant	lda, qda	<i>MASS</i>
	rforest	<i>randomForest</i>	analysis	fda	<i>mda</i>
MARS	mars	<i>mda</i>	clustering	hclust	<i>mva</i>
	polymars	<i>polymars</i>		hierclust	<i>multiv</i>
				kmeans	<i>mva</i>
PPR	PPR	<i>modreg</i>	convex	cclust	<i>cclust</i>
ridge regression	lm.ridge	<i>MASS</i>	clustering		
linear regression	lm	<i>(base)</i>	fuzzy	fanny	<i>cluster</i>
robust regression	lqs	<i>lqs</i>	clustering	cmeans	<i>e1071</i>
polyn. regression	polyreg	<i>mda</i>	EM algorithm	emclust	<i>mclust</i>
GLM, logistic regression	glm	<i>(base)</i>	polyclass	polyclass	<i>polyspline</i>
GAM	gam	<i>mgcv</i>	Logit-/Adaboost	gbm	<i>gbm</i>
non-linear regression	loess	<i>modreg</i>	Bayesian	learn	<i>deal</i>
	gls, glms	<i>nlme</i>	Network Learning		
	optim	<i>(base)</i>			



R and WinEdt as Development Environment



```
R-WinEdt - [C:\home\hwb\Programming\Rwin\code\outlier-mad.R]
File Edit Format Search Insert Tools Options Window Help R
outlier-mad.R
#-----
#
# outlier.mad
#
#
outlier.mad <- function (x, k)
{
  # x: vector or time series
  # k: window [x_(i-k),...,x_i,...,x_(i+k)]
  n <- length(x)
  y <- x # corrected x vector
  ind <- c() # indices of outliers

  L <- 1.4826 # constants for normal distributions
  t <- 3 # Pearson's 3 sigma edit rule

  # we don't look at outliers at the end parts of x !
  for ( i in (k+1):(n-k) ) {
    x0 <- median( x[(i-k):(i+k)] )
    S0 <- L * median( abs(x[(i-k):(i+k)] - x0) )
    if ( abs(x[i]-x0) > t * S0 ) {
      y[i] <- x0
      ind <- c(ind, i)
    }
  }
  # return a list with 2 components
  list(y=y, ind=ind)
}
```

WinEdt...

- LaTeX and ASCII text editor
- ...as an **IDE for R**
- syntax highlighting for the S language
- know-how about **R** functions
- interfacing to **R**, i. e. running scripts or selected code pieces

(X)Emacs...

- ESS package



The (D)COM Interface for R



Integration with Excel

The screenshot shows the REExcel menu with options: R Start, Run Code, Get, Put, Set R working dir, Copy Code, Debug R, Options, REExcel Help, R Help, and About REExcel. Below the menu is a table of R functions:

	A	B	C	D
1	R server connection demo			Imp
2	Sheet functions for Rserver			
3				
4	The next 2 sheets give examples			Eric
5	of the R sheet functions			A
6				
7	These are the functions:			
8				
9	RVarSet		assigns a scalar R expression to an R variable	
10	RPut		assigns a scalar (numeric) value from an Excel cell	
11	RStrPut		assigns a scalar (string) value from an Excel cell	
12				
13	REval		evaluates an R expression with dependencies	
14	RApply		applies a function to arguments	
15			These function may be used as array functions	
16				
17	RProc		runs a sequence of R commands	
18				
	mat1			
	mat1			=RVarSet(A7,"diag(c(11:13))")
	11	0	0	=REval(A8)
	0	12	0	
	0	0	13	

Integration with Visual Basic

VBA macros for accessing R can be written using the following procedures and functions:

```

RInterface.StartRServer()
    Starts the R server
RInterface.StopRServer()
    Stops the R server
RInterface.RRun(commandstring)
    Executes commandstring
RInterface.PutArray(varname,range)
    Assigns the contents of range to R variable var
RInterface.GetArray(varname,range)
    Puts the value of R variable var into Excel range range
    
```

Accessing R from Visual Basic:

```

dim x as StatConnector
set x = new StatConnector
on error goto error_handler

x.Init ("R")

x.SetSymbol ("symname",value)
' y = x.GetSymbol ("symname")

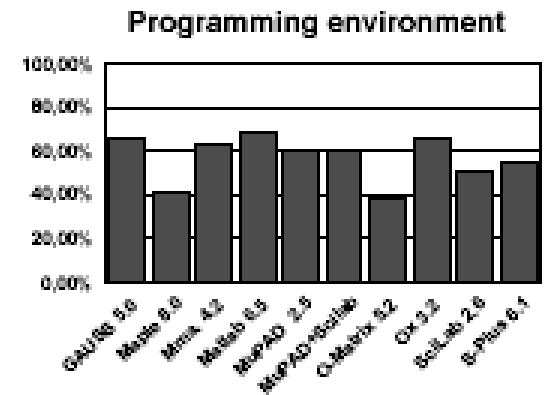
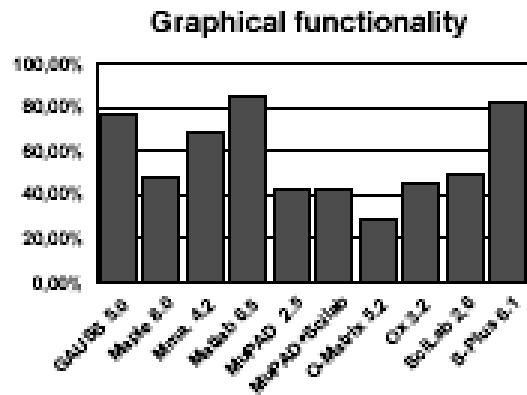
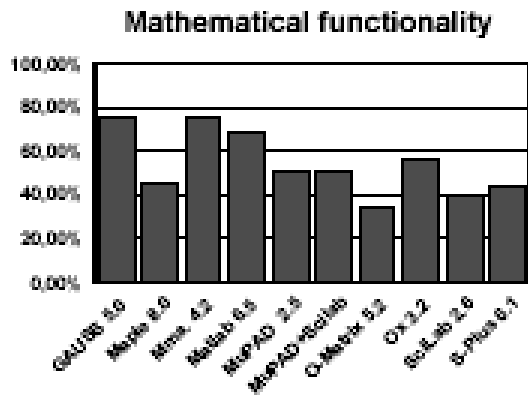
y = x.Evaluate ("expression")
' x.EvaluateNoReturn ("expression")

x.Close

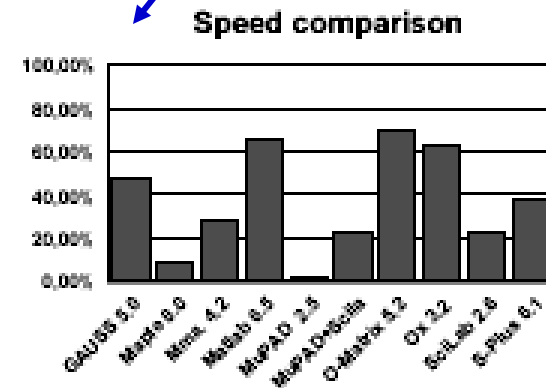
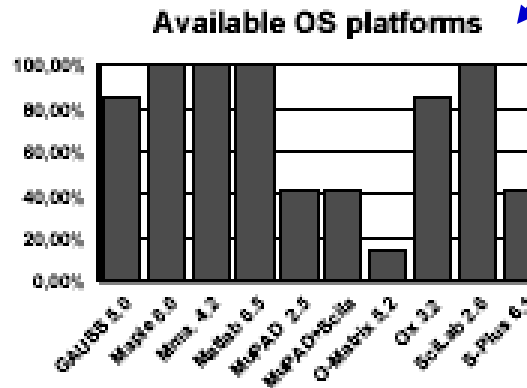
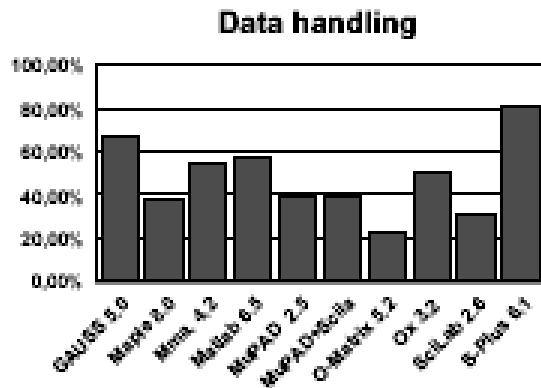
error_handler:
MsgBox x.GetErrorText,"R Server Error"
    
```



Compare Programs for Data Analysis



Roughly, use S-Plus 6.1 as comparison for R;
but: R is (much) better than S-Plus in platforms and speed !



Source: St. Steinhaus, University of Frankfurt, 2002
<http://www.scientificweb.de/ncrunch/>



References

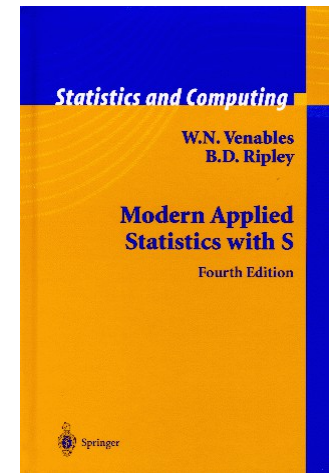


■ Internet Pages

- R Home <http://www.r-project.org/> (TU Wien)
- CRAN <http://stat.ethz.ch/CRAN/> (ETH Zürich)

■ Books, for instance

- W. N. Venables and **B. D. Ripley** (2003)
Modern Applied Statistics with S (4th Edition)



■ On-line Documentation

- **R News:** The Newsletter of the R Project
- **R Mailing List:** 'r-help', 'r-package', 'r-develop', and 'r-announce'
 - 30-40 entries in 'r-help' every day (Jan. – May 2003)

■ Related Projects:

ADE4, Bioconductor, Omega, gRaphical Models, and R GUIs



ABB