A few resources

(Connecting to the seminar web site)

This will remain available to everyone
R statistical software

About R
What is R?
Contributors
Screenshots
What's new?

Download, Packages
CRAN

R Project
Foundation
Members & Donors
Mailing Lists
Bug Tracking
Developer Page
Conferences
Search

Documentation
 Manuals
 FAQs
The R Journal
 Wiki
 Books
 Certification
 Other

Misc
Bioconductor
Related Projects
User Groups
Links

Getting Started:
- R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred CRAN mirror.
- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News:
- R 2.11.0 prerelease versions will appear starting March 25. Final release is scheduled for April 22, 2010.
- R version 2.10.1 has been released on 2009-12-14. The source code will first become available in this directory, and eventually via all of CRAN. Binaries will arrive in due course (see download instructions above).
- The first issue of The R Journal is now available
- useR! 2010, the R user conference, will be held at NIST, Gaithersburg, Maryland, USA, July 21-23, 2010.
Why is it popular?

- Free *and* professional-quality standard
- Open source
  - Everyone can see what the program does
  - Everyone can contribute
  - Everyone can use it, discuss it, criticize it
- Ideal for teaching, especially under reduced budgets
R object types

- **Vector:**
  - a one-dimensional array of arbitrary length. Subsets of the vector may be referenced. All elements of the vector must be of the same type--numerical, character, etc.

- **Matrix:**
  - a two-dimensional array with an arbitrary number of rows and columns. Subsets of the matrix may be referenced, and individual rows and columns of the matrix may be handled as vectors. Again all elements of the matrix must be of the same type.

- **Array:**
  - as a matrix, but of arbitrary dimension (can be greater than 2-D).

- **Data frame:**
  - a set of data organized similarly to a matrix. However each column of the data frame may contain its own type of data. Columns typically correspond to variables in a statistical study, while rows correspond to observations of these variables. A data frame may be handled similarly to a matrix, and individual columns of the data frame may be handled as vectors.

- **List:**
  - an arbitrary collection of other R objects (which may include other lists).

- **Function:**
  - A set of R commands...too advanced for now
Main presentation topics

- A few important advantages
- How to import and export data
- Graphical displays
- Data manipulation
- Examples of more advanced analyses
- Tree-ring applications

(all very brief, since we have little time, but the purpose is to get you started)
Data import

- Easy to do if you know how...
- Specific commands exist for different formats
- Download the file from the web site and import it into R
R commands can be saved

• Notice how I am using R commands that were saved in a file (so that I could re-issue them today)

• Anyone can program!
  – Students learn even if they are math-phobics

• Extremely useful feature for long-distance communication, work revisions, research archiving and sharing
  – Improves transparency and repeatability
Box-and-whisker plots

Example Data Set

Site ID

PPO

PNA

CAN

Chronology Value (Mean Ring Index)
Histogram => Data distribution

Is this a normally distributed data set?
Samples vs. populations

• In tree-ring science we typically deal with samples that are not taken “at random”, rather they are observations, without any treatment or control

• Even “random” samples can differ quite a bit from the underlying theoretical model

• R allows for visualizing and fitting a number of probability distributions
Theoretical distributions in R

• Many different types (see handout on web site)

• Discrete vs. continuous variables

• The “normal” (or Gaussian) distribution is not the only one to be bell-shaped
A “random” sample

Histogram of the Z random sample

Sample values (n=100)
The linear regression model

\[ y_i = \alpha + \beta x_i + \varepsilon_i \]

\[ = \hat{y}_i + e_i \]

- The dependent (or response) variable \( y \) (also called “predictand”) is a function of the independent (or explanatory) variable \( x \) (also called “predictor”).

- The intercept and slope are the parameters of the model
“Least squares” estimation

If \( \hat{y}_i = a + bx_i \)

and \( e_i = y_i - \hat{y}_i \)

then \( \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \)

\[ = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 \]

We can then calculate \( a \) and \( b \) (the estimates of \( \alpha \) and \( \beta \)) so that their values minimize this sum of squares (= residual SS = unexplained SS)
Always plot the data!

Four OLS regression charts with the same $R^2$, slope, and intercept.

Figure 9.18  Three key pathologies in regression (after Anscombe, 1973).  
Example of simple linear regression
Okay, now listen up. Nobody gets in here without answering the following question: A train leaves Philadelphia at 1:00 p.m. It’s traveling at 65 miles per hour. Another train leaves Denver at 4:00... Say, you need some paper?

Math phobic’s nightmare
Multivariate Analysis

Multidimensional data-space

Each measurement is represented by a vector (pattern) in a proper N-Dim space (patterns space).

\( \mu \) is the mean vector the variance (\( \sigma \)) is replaced by the covariance matrix (\( \Sigma \)).

\[
\mu = \begin{bmatrix} \mu_1 & \cdots & \mu_d \end{bmatrix}
\]

\[
\Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1d} \\
\vdots & \ddots & \vdots \\
\sigma_{d1} & \cdots & \sigma_{dd} \end{bmatrix}
\]

\( S = \{S_1, \ldots, S_n\} \)
**PCA: Geometric Interpretation**

**Definition**
Let us consider a matrix $X$ of data, let $C = X^TX$ be the correlation matrix of $X$.
The $i$-th principal component of $X$ is $X^Tc(i)$, where $c(i)$ is the $i$-th normalized eigenvector of $C$ corresponding to the $i$-th largest eigenvalue.
Eigenvalues in R

|                  | PC1   | PC2   | PC3   | PC4   | PC5   | PC6   | PC7   | PC8   | PC9   | PC10  | PC11  | PC12  | PC13  | PC14  | PC15  | PC16  | PC17  | PC18  | PC19  | PC20  | PC21  | PC22  | PC23  | PC24  |
|------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Standard deviation| 3.229 | 1.783 | 1.324 | 1.138 | 1.1218| 0.9701| 0.9344| 0.7991| 0.7499| 0.6966| 0.6635| 0.6095| 0.5744| 0.5214| 0.4987| 0.44292| 0.42627| 0.37226| 0.32903| 0.28457| 0.27583| 0.24918| 0.2137| 0.18267 |
| Proportion of Variance | 0.434 | 0.133 | 0.073 | 0.054 | 0.0524| 0.0392| 0.0364| 0.0266| 0.0234| 0.0202| 0.0184| 0.0155| 0.0138| 0.0113| 0.0104| 0.00817| 0.00757| 0.00577| 0.00451| 0.00337| 0.00317| 0.00259| 0.0019 | 0.00139  |
| Cumulative Proportion | 0.434 | 0.567 | 0.640 | 0.694 | 0.7464| 0.7856| 0.8220| 0.8486| 0.8721| 0.8923| 0.9106| 0.9261| 0.9399| 0.9512| 0.9616| 0.96972| 0.97729| 0.98306| 0.98758| 0.99095| 0.99412| 0.99671| 0.9986 | 1.00000  |

Importance of components:
Plot of the main PCs (sorted by eigenvalue variance)
First three Eigen-vectors

First 3 PCS of Tree-Ring Chronologies

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAD</td>
<td>-0.215191928</td>
<td>-0.016359057</td>
<td>0.094070138</td>
</tr>
<tr>
<td>BSA</td>
<td>-0.222770195</td>
<td>-0.206989385</td>
<td>0.149046275</td>
</tr>
<tr>
<td>CAN</td>
<td>0.004094265</td>
<td>-0.208639169</td>
<td>0.595119832</td>
</tr>
<tr>
<td>CIM</td>
<td>-0.237926144</td>
<td>0.151709748</td>
<td>0.213283642</td>
</tr>
<tr>
<td>FOG</td>
<td>-0.156828867</td>
<td>0.230187316</td>
<td>0.326500340</td>
</tr>
<tr>
<td>FOU</td>
<td>-0.244363477</td>
<td>0.137099910</td>
<td>-0.156819922</td>
</tr>
<tr>
<td>LEP</td>
<td>-0.239959794</td>
<td>0.188032278</td>
<td>-0.006115967</td>
</tr>
<tr>
<td>MAI</td>
<td>-0.249627651</td>
<td>0.004910725</td>
<td>0.007734328</td>
</tr>
<tr>
<td>MBA</td>
<td>-0.203290439</td>
<td>0.174919909</td>
<td>0.005560333</td>
</tr>
<tr>
<td>MOR</td>
<td>-0.240423697</td>
<td>-0.055188590</td>
<td>0.159680443</td>
</tr>
<tr>
<td>MSC</td>
<td>-0.163735808</td>
<td>0.054358789</td>
<td>0.007734328</td>
</tr>
<tr>
<td>ORI</td>
<td>-0.094265619</td>
<td>0.204898150</td>
<td>0.007734328</td>
</tr>
<tr>
<td>PET</td>
<td>-0.203489815</td>
<td>0.004910725</td>
<td>0.007734328</td>
</tr>
<tr>
<td>PIZ</td>
<td>-0.252807455</td>
<td>-0.055188590</td>
<td>0.159680443</td>
</tr>
<tr>
<td>PNA</td>
<td>-0.202730549</td>
<td>0.174919909</td>
<td>0.005560333</td>
</tr>
<tr>
<td>PPO</td>
<td>-0.142931778</td>
<td>0.004910725</td>
<td>0.007734328</td>
</tr>
<tr>
<td>ROS</td>
<td>-0.248916309</td>
<td>0.004910725</td>
<td>0.007734328</td>
</tr>
<tr>
<td>SAB</td>
<td>-0.243113778</td>
<td>0.004910725</td>
<td>0.007734328</td>
</tr>
<tr>
<td>TER</td>
<td>-0.239824125</td>
<td>0.004910725</td>
<td>0.007734328</td>
</tr>
<tr>
<td>TOC</td>
<td>-0.128392734</td>
<td>0.004910725</td>
<td>0.007734328</td>
</tr>
<tr>
<td>VCH</td>
<td>-0.206507241</td>
<td>0.004910725</td>
<td>0.007734328</td>
</tr>
<tr>
<td>VCL</td>
<td>-0.226106374</td>
<td>0.004910725</td>
<td>0.007734328</td>
</tr>
<tr>
<td>VAL</td>
<td>-0.132110275</td>
<td>0.004910725</td>
<td>0.007734328</td>
</tr>
<tr>
<td>VEN</td>
<td>-0.187817811</td>
<td>0.004910725</td>
<td>0.007734328</td>
</tr>
</tbody>
</table>
PC Scores or Amplitudes

Reversed scores (red) and mean values (black)

Remember to check the overall sign!!!
dp(IR)-specific commands

• We’ll go through the code in R

• The resulting graphs are available as a poster, and I hope you will stop by to take a look at it...Thanks for your time!
**Dendrochronology in R with the dplR library**

**Abstract**

We demonstrate the utility of the dendrochronological program library in R (dplR; Bunn 2008) for performing standard tree-ring analyses including crossdating, detrending, and chronology building. R is a well-regarded open-source statistical computing environment where users can contribute packages that are freely available. The detrending options in dplR have been updated and now include regional curve and C-method standardization (Biondi and Qeadan 2008). We also demonstrate crossdating in dplR via moving correlation analyses, cross correlation, and automatically generated skeleton plots. The expansion of dplR is largely due to code contributed, in part or in whole, by the user community. The increased functionality of dplR makes it easier for dendrochronologists to use R as their primary analytic environment.

**Data display and crossdating**

Ring-width data used to illustrate the use of dplR are from the Gus Pearson Natural Area, in Northern Arizona, USA. This area has been monitored since the early 1900s for changes in forest structure, while at the same time being protected from logging or wildfire. Changes in spatio-temporal patterns of tree density, survival, and stem growth were documented using repeated forest inventories, whose trends were placed in a longer historical context using dendrochronological analysis (for additional details, see Biondi 1996, 1999).

Dendrochronology in R with the dplR library
Andrew G. Bunn and Franco Biondi

**Standardization and chronology development**

Using dplR and other R code, it is easy to visualize the effect of many standardization options. First, as a sort of “control,” one can plot the median of the ring-width series (shown on the left). Then, standardization can be obtained by computing ring areas (~ annual basal area increment; Biondi 1999) going either from the pith towards the bark (inside-out) or in the opposite direction (outside-in). The median tree-ring chronologies from ring areas computed in these two ways are shown on the right.

**Available commands in the current version of the dplR library are shown below.**

**Other options**

**References**


**Acknowledgments**

We thank all the people who generously provided code for inclusion in dplR, particularly Mike Forster and Christian Zang. Additional R code was provided by Christof Bigler and Björn Reineking. F. Biondi was partially supported by NSF Cooperative Agreement EPS-0814372 to the Nevada System of Higher Education.

**Other crossdating tools are moving window correlations (left), time-series and scatter-plots (right). Additional details are given by Bunn (in press).**

**Ring-width data used to illustrate the use of dplR are from the Gus Pearson Natural Area, in Northern Arizona, USA. This area has been monitored since the early 1900s for changes in forest structure, while at the same time being protected from logging or wildfire. Changes in spatio-temporal patterns of tree density, survival, and stem growth were documented using repeated forest inventories, whose trends were placed in a longer historical context using dendrochronological analysis (for additional details, see Biondi 1996, 1999).**

**Dendrochronology in R with the dplR library**

Andrew G. Bunn and Franco Biondi

**Data display and crossdating**

Ring-width data used to illustrate the use of dplR are from the Gus Pearson Natural Area, in Northern Arizona, USA. This area has been monitored since the early 1900s for changes in forest structure, while at the same time being protected from logging or wildfire. Changes in spatio-temporal patterns of tree density, survival, and stem growth were documented using repeated forest inventories, whose trends were placed in a longer historical context using dendrochronological analysis (for additional details, see Biondi 1996, 1999).