ESSReS-L2:

Part I: 16 Feb – 20 Feb, 2009

Part II: 11 May – 12 May, 2009

Introduction to the interdisciplinary field of Earth System Science Research, Part II

"Introduction to computational techniques and statistical data analysis"

Editor in charge: Klaus Grosfeld,
Earth System Science Research School, 2009
Alfred Wegener Institute, Bremerhaven
info@earth-system-science.org

Course programme within the ESSReS Curriculum

ESSReS-L2: Introduction to the interdisciplinary field of Earth System Science Research, Part II

Introduction to computational techniques and statistical data analyses
Block course: Part I: 16 - 20 February 2009, one week, 9 am - 5 pm
Part II: 11 – 12 May 2009, two days, 9 am – 5 pm
Location: Jacobs University , AWI
*Responsible: **V. Unnithan, S. Frickenhaus**, **M. Mudelsee**, P. Baumann,, A. Gelessus, H. Grobe, T. Laepple, L. Linsen, A. Schaefer, R. Sieger*
Email: info@earth-system-science.org
v.unnithan@jacobs-university.de
Stephan.Frickenhaus@awi.de
Manfred.Mudelsee@awi.de

A better understanding of the Earth System requires advanced methods and techniques in data analysis and modelling. This can be gained by statistical data analysis of time series as well as complex numerical models. With this course, training and education in different computational methods/platforms as well as data analysis techniques is fostered as key components, when investigating local processes in a global context.

Part I:

Unit 1: **Date: Monday, 16 February 2009**

Location: Jacobs University

Building: West Hall

Room: *CLAMV basement*

9:00 – 10:30: Introduction to UNIX (Achim Gelessus, V. Unnithan)

11:00 – 12:30: Introduction to visualization methods (Lars Linsen)

14:00 – 15:30: Visualization methods and tool presentation (Lars Linsen)

15:30 – 16:30: Introduction to GIS: Attribute data, values and classification (Angela Schäfer)

Unit 2: **Date: Tuesday, 17 February 2009**

Location: Jacobs University

Building: West Hall

Room: *CLAMV basement*

9:00 – 10:30: Data banking and webifying (Peter Baumann)

11:00 – 12:30: continued

14:00 – 15:30: continued, max until 17:00

Unit 3: **Date: Wednesday, 18 February 2009**

Location: AWI

Building: F

Room: *Glaskasten*

9:00 – 10:30: Archiving of data from earth system research – an overview (H. Grobe, R. Sieger)

11:00 – 12:30: Data archiving, compilation and visualization – a show case (R. Sieger, H. Grobe)

Unit 4: **Date: Thursday, 19 February 2009**

Location: AWI

Building: F

Room: *Glaskasten*

14:00 – 17:00: Introduction to 'R' as a statistical analysis platform, Part I (S. Frickenhaus)

Unit 5: **Date: Friday, 10 October 2008**

Location: AWI

Building: F

Room: *Glaskasten*

14:00 – 17:00: Introduction to 'R' as a statistical analysis platform, Part II (S. Frickenhaus)

Part II:

Unit 6: **Date: Monday, 11 May 2009**

Location: AWI

Building: F

Room: *Glaskasten*

10:00 – 11:30 Statistical Sciences: regression, distrubution functions, persistence (M. Muddelsee)

12:00 – 14:00: Recapitulation of 'R', statistical practices (T. Laepple)

Unit 7: **Date: Tuesday, 12 May 2009**

Location: AWI

Building: F

Room: *Glaskasten*

9:00 – 10:30: Bootstrap: linear regression II, Monte Carlo Experiments, Ramp (M. Muddelsee)

11:00 – 12:30: Statistical practices (T. Laepple)

13:30 – 15:00: Lecture continued (M. Muddelsee)

15:30 – 17:30: Statistical practices (T. Laepple)

Updates of the course program and location can be found at

http://www.earth-system-science.org/en/courses/

Contents: page

Data banking and webifying

Lecturer
Name: Peter Baumann
Department: EECS
Institute: Jacobs University Bremen
Email: p.baumann@jacobs-university.de
Duration of lecture: 4 lectures á 90 minutes (including practices)

Lecture content

This course unit will introduce you to setting up an exemplary geo Web service using standard, open-source Web and database technology. The course will give a brief, practically oriented overview on concepts and then dive into a mini project to set up a small working web service encompassing meta, vector, and raster data.

By the end of this unit you will know how meta, vector, and raster data can be retrieved from geo databases and how they can be connected in a service. You will be able to build your own small web app using PHP and PostgreSQL.

Useful topics to look at in preparation of this unit:

- HTML (up to the level of writing simple pages yourself, that is: without an HTML editing tool)
- Linux command line, such as ls; pwd; cd
- some Linux-available editor, such as vi or emacs
- SQL

optional:

- PHP
- querying PostGIS databases

Introduction to ‘R’ as a statistics analysis platform

Lecturer
Name: Stephan Frickenhaus
Department: Biosciences
Institute: AWI
Email: stephan.frickenhaus@awi.de
Duration of lecture: 2 lectures á 60 minutes (including practices)

Lecture content

R is a scripting language with several advantages over other software products to develop data analysis pipelines suitable for the preparation and documentation of research articles.

The course is demonstrating the interactive features of R and some of ist extension packages. Basic statistic analysis principles are reviewed, like the concept of confidence interval and p-value for hypothesis testing. The power of R in data modeling within linear models is shown. Classification and clustering techniques, including PCA and hierarchical clustering, are introduced, based on practical approaches to handle sample data.

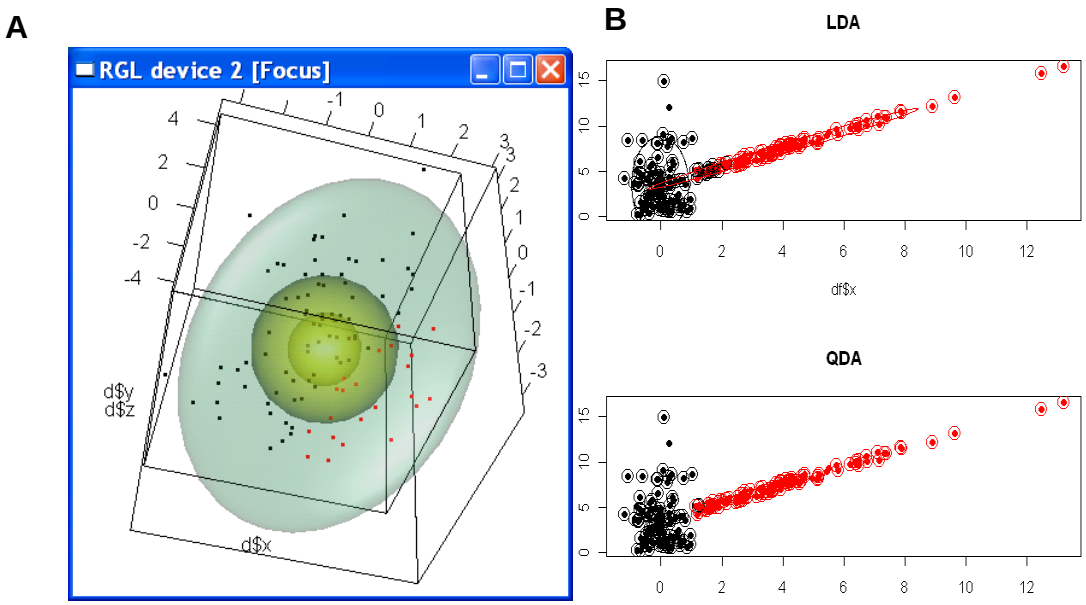


Figure: **A**) An interactive 3D-view of two groups of data (red, black) within their covariance ellipsoid (green) and two spheres (yellow). **B**) Linear (top) vs. Quadratic (bottom) Discrimination Analysis of two groups of data (red, black). False classified data indicated by black circles with red disks inside.

The course starts teaching the R basics of data-representation in objects like lists, vectors, matrices and data-frames, visualisation by scatter-plots, box-plots and covariance ellipses, simple tests and advanced analysis methods like ANOVA. The second part of the course is focuses clustering algorithms and the idea of machine learning. Besides statistics and R, principles and advantages of the OpenAccess and OpenSource are discussed as paradigms enabling a higher level of transparency in scientific research processes and publication practice.

References

R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.r-project.org.

Statistics: An Introduction Using R (2005), Michael J. Crawley, Wiley & Sons

Introduction to UNIX/LINUX

Lecturer
Name: Achim Gelessus
Department: CLAMV
Institute: Jacobs University
Email: a.gelessus@jacobs-university.de de
Duration of lecture: 150 min

Lecture content

This course provides an introduction to the UNIX operating system. UNIX in all its many flavours is a widely in the scientific community used computer operating system. UNIX is a resource sharing operating system with multi-user and multi-tasking capabilities. Early versions of UNIX were released in 1969. UNIX has gone through a long process of modifications and improvements since then. The porting of UNIX to standard PC hardware in the 1990's is called Linux and in the meantime Linux has replaced many of the older UNIX versions.
The lecture gives an introduction to the UNIX file system including file permissions. The most important commands for UNIX users are discussed and explained by examples. Further topics are redirections, pipes, vi editor and remote access.

Literature

- Introduction to UNIX and Linux: Tutorial lectures and exercise sheets
  http://www.doc.ic.ac.uk/~wjk/UnixIntro/index.html

- Ellen Siever, Stephen Figins, Aaron Weber, Lars Schulten
  Linux in a Nutshell
  O'Reilly 2005

- Daniel J. Barrett
  Linux Pocket Guide
  O'Reilly 2004

- Dave Taylor
  Sams Teach Yourself Unix in 24 Hours
  Sams 2005

- Amir Afzal
  UNIX Unbounded
  Prentice Hall 2007

- David I. Schwartz
  Introduction to Unix
  Prentice Hall 2005

- Linus Torvalds
  Just For Fun – The Story of an Accidental Revolutionary
  HarperCollins 2001

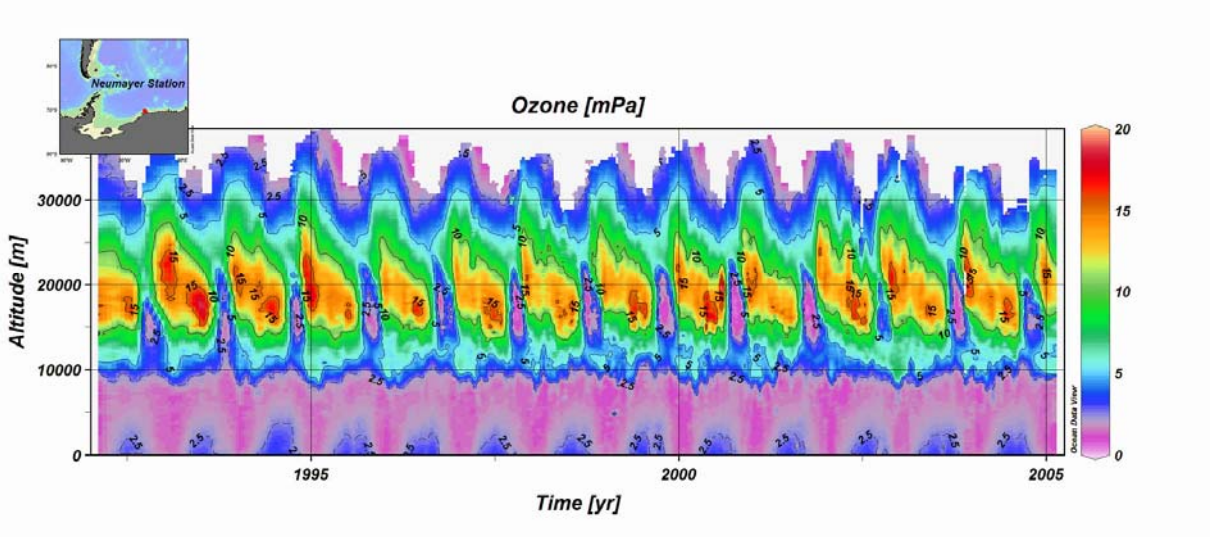
I) Archiving of data from earth system research - an overview
II) Data archiving, compilation and visualization - a show case

Lecturer
Name: Hannes Grobe, Rainer Sieger
Department: Geoscience
Institute: AWI
Email: Hannes.Grobe@awi.de, Rainer.Sieger@awi.de
Duration of lecture: 2 times 1.5 hours

Lecture content

Unit I: With examples from earth system research data, an overview is presented on organizations, infrastructures and technologies involved in the long-term archiving, publication and provision of data, e.g. Global Change Master Directory, World Data Center System, NASA, national archives, portals. The importance of metadata, geocodes, citations and persistent identifiers in the context of data search and usage is discussed. Major international projects producing data and repositories storing data in the field of earth sciences are presented in terms of the 5 earth spheres (hydrosphere, cryosphere, atmosphere, lithosphere, biosphere).

Unit II: The workflow in data archiving is explained by using the information system PANGAEA© as an example. PANGAEA is an information system, operated as Open Access library for geo-referenced data from basic research on earth and environment. Scientific primary data are archived with related metainformation and are distributed via web-services and accessible through various clients. An introduction is given on how to find data, how to compile larger data collections and finally how to use visualization software to present those data. The required technology in the background is briefly explained (relational database, data warehouse, backup).



Compilation of an Ozone measurement time series from the Antarctic (König-Langlo & Gernandt 2009)

Literature/Web-links

World Data Center System (ICSU)	http://www.ngdc.noaa.gov/wdc
WDC Portal (beta)	http://www.world-data-system.org
International Council for the Exploration of the Sea (ICES)	http://www.ices.dk
Ocean Biogeographic Information System	http://www.iobis.org
Global Biodiversity Information Facility	http://www.gbif.org
Software *Ocean Data View*	http://odv.awi.de
eWOCE - Electronic Atlas of WOCE Data	http://www.ewoce.org
PanPlot & Pan2Applic	http://www.pangaea.de/software
PANGAEA Publishing Network for Geoscientific & Environmental Data	http://www.pangaea.de
Diepenbroek et al. 2002 in Computers & Geosciences	doi:10.1016/S0098-3004(02)00039-0
Earth System Science Date - The Data Publishing Journal	http://www.earth-system-science-data.net
OAIster (portal for digital resources)	http://www.oaister.org
ScientificCommons (beta, public portal for publications s.l.)	http://www.scientificcommons.org
ScienceDirect (portal of Elsevier)	http://www.sciencedirect.com

The impact of visualization

Lecturer
Name: Lars Linsen
Department: Computational Science & Computer Science
Institute: Jacobs University
Email: l.linsen@jacobs-university.de
Duration of lecture: 2.5 hours (including 0.5 hours of tool presentation)

Lecture content

Scientific visualization deals with the visualization of data with spatial interpretation such as computer-generated data from numerical simulations or measured data using scanning or sensoring techniques. The goal of scientific visualization is to extract salient features from the given data and represent them visually such that an intuitive and correct understanding of the extracted feature is possible. Instead of a static display of the feature, interactive operations and animations support dynamic views on the data allowing for an interactive visual exploration of the data. This lecture will provide an introduction to methods for visualization using 2D and 3D displaying techniques. Standard algorithms and concepts will be taught. The ideas presented in the lecture on “the impact of visualization” will be elaborated and presented in a methodological fashion. Towards the end of the lecture, some tools for visualization of geospatial data will be presented.

Literature

Alexandru Telea: *Data Visualization: Principles and Practice*, Wellesley, Mass.: AK Peters, 1st edition, 2008.
Charles D. Hansen & Christopher R. Johnson: *Handbook of Visualization*. Academic Press, 2004.

Statistical climate data analysis

Lecturer
Name: Manfred Mudelsee
Department: Climate Science/ Paleodynamics
Institute: AWI
Email: mudelsee@mudelsee.com
Duration of lecture: 3 times 1.5 hours

Lecture content

The major role of statistical analysis for the applied sciences is to provide error bars (confidence intervals, etc.) for estimations of parameters. Estimates without error bars are useless. The difficulty with climate data is that assumptions often made in conventional analysis techniques (Gaussian distributional shape, absence of persistence) usually fail here. Hence, more advanced methods are required.

The lectures start with an introduction to conventional estimation techniques and the "statistical language" and notation. We explore then a simple estimation problem: mean levels of carbon dioxide concentrations during glacials and interglacials. Next is linear regression as a tool to quantify trends in climate time series. The bootstrap is introduced as a powerful tool to obtain realistic error bars in the "non-conventional" climate situation.

Finally, we turn to nonlinear ramp function regression (Mudelsee 2000) as a method to measure climate transitions. We analyse Dansgaard-Oeschger event 5 on various proxy variables in the NGRIP ice core (see Figure).

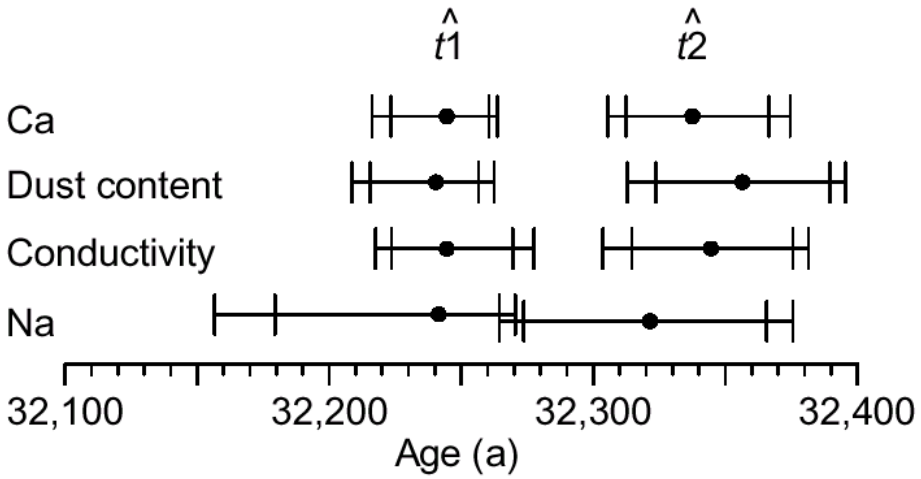


Figure 4.15. Onset of Dansgaard–Oeschger event 5, NGRIP ice core: estimated change-points with confidence intervals. Shown are 95% and 90% BCa CIs for $\widehat{t1}$ and $\widehat{t2}$, calculated with ARB resampling.

Literature

Mudelsee M (2000) Ramp function regression: A tool for quantifying climate transitions. *Computers and Geosciences* 26:293–307

Mudelsee M (in prep.) Climate Time Series Analysis: Classical and Bootstrap Methods. Springer [copy of Chapters 1 to 4 provided for ESSReS students]

von Storch H, Zwiers FW (1999) Statistical Analysis in Climate Research. Cambridge Univ. Press.

Statistical climate data analysis in ‘R’

Lecturer
Name: Thomas Laepple
Department: Climate Science/ Paleodynamics
Institute: AWI
Email: tlaepple@awi.com
Duration of lecture: 4 times 1.5 hours

Lecture content

This lecture introduces basic statistical climate data analysis using the scripting language R with focus on bootstrap methods.

In hands-on experiments, we first recapitulate the basics of the R-programming language. Basic statistical analysis as linear regression and the role of confidence intervals are introduced.
The second day covers the use of ordinary bootstrap, as well as more advanced bootstrapping methods. In exercises on artificial and real data, the use of these methods, and the importance of taking into the correct assumptions (persistence, distributional shape) are demonstrated.

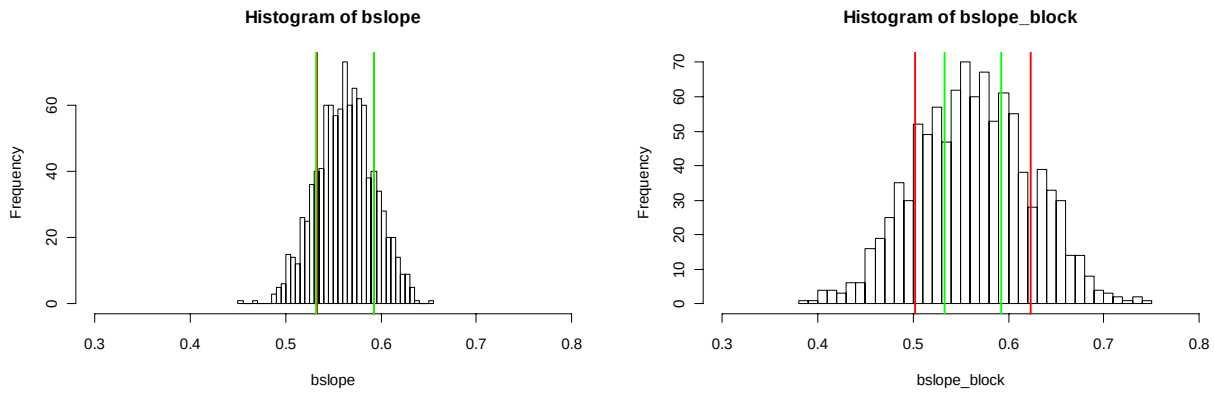


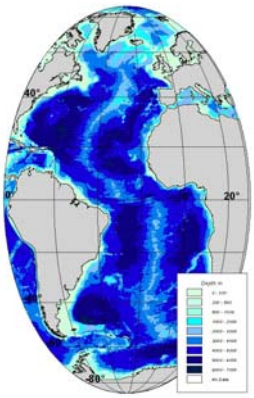
Figure 1: Bootstrap distribution of the global mean temperature trend of the last century. Left panel: Ordinary bootstrap; Right panel: Block bootstrap, taking into account the temporal persistence of the temperature data. Additionally the classical standard error (green) and the bootstrap standard error (red) are shown. This example shows the importance of taking the serial dependence into account when estimating confidence intervals.

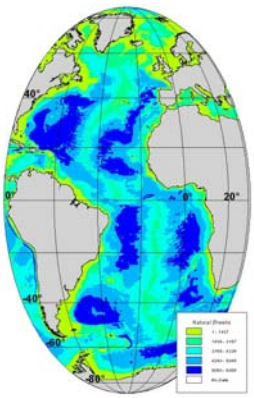
Literature

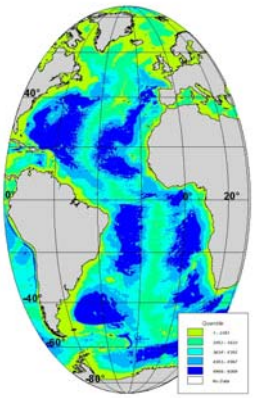
www.r-project.org

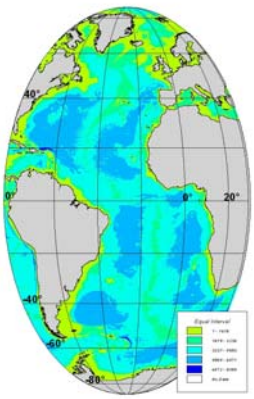
Visualisation techniques for spatial data and statistical classification methods

Lecturer
Name: Angela Schaefer
Department: MARUM
Institute: University of Bremen
Email: a.schaefer@uni-bremen.de
Duration of lecture: 90 min









Lecture content

In earth sciences modelling of spatial phenomenon and resulting visualisation by means of thematic maps is a common but not trivial major task. When you present spatial data in form of thematic map models a graphical presentation of geographic data is created. To be effective, this map must be visually compelling and thematically convincing. Hence a map is the interface between a spatial data model and our perception. Maps utilize people's inherent cognitive abilities to identify spatial patterns and provide visual cues about the qualities of the visualized objects and locations. Since a map is always an abstraction of your data it filters information for the intended purpose. Therefore geo-statistical concepts and rules must be considered to simplify data and to blend out some of the complexity or internal structure of data.

This lecture provides an introduction into adding descriptive content to spatial data models — categories, types, symbols, labels, and other visual information.

- Drawing selected thematic views of spatial data
- Techniques for illustrating measured, classified and descriptive attributes of data
- Measured and descriptive values and what they can represent: ration, interval, ordinal, nominal.
- Types of attributes and application: float, double, integer, identifiers, text, date, BLOBs
- Types of data represented in raster cells: discrete and continuous raster data
- Standard classification scheme and application examples: equal interval, quantile, smart quantiles, natural breaks, defined intervals

Literature

Zeiler, M. (1999). Modeling Our World: The ESRI Guide to Geodatabase Design. Redlands, Calif., ESRI press. http://books.google.de/books?id=qAe-ScoyTqIC