Methods and Analysis of Large Data Sets:
DATA! DATA! DATA!

Mon-Thur, 9:00-10:50, Summer Term A, 2014.

Stephan Sain and Doug Nychka
Affiliate Faculty, Applied Mathematics
Scientific Staff, National Center for Atmospheric Research

Course Goals

Statistics is the science of interpreting data through mathematical models with an emphasis on quantifying the uncertainty in any analysis. Typically part of a statistical analysis will also involve the use of graphics to communicate complex relationships and patterns. Traditional courses in statistics focus on developing the mathematical basis of statistical concepts, for example sampling from a population, using probability models for testing hypotheses and setting error bounds for parameter estimates. However, this view may miss the rich set of tools that can be applied to different kinds of data. Moreover, the rapid increase in computing power and the R statistical language has made statistical methods accessible to a broad scientific and engineering community and with it the ability to interpret large and complex data sets. It is the goal of this course to focus on the power of statistical methods to deal with data that would be difficult to interpret using elementary statistics and limited graphics. Some overall themes are being able to recognize different types of data and statistical questions and being able to identify statistical tools that are appropriate. We will take the approach that many sophisticated and advanced methods can be appreciated and used within the context of particular data sets if students have a clear ideas of the analysis goals and an understanding of how the data is collected or generated. This kind of understanding is a practical complement to the more mathematical development that would occur in other courses.

Class will be taught the R language for data analysis and will become adept in its use.

Outline

June 2-5 Exploratory data analysis

M R Bootcamp
Tu: Programming in R
W: Tools to summarize and visualize data
Th: Tools to summarize and visualize data (con't)

June 9-12 Modeling relationships within a dataset

M Fitting distributions
Tu: Regression

W: Regression
Th: Smoothers: Fitting curves and surfaces to data.

June 16-19 Modeling dependence within a data set
M Times Series
T: Times Series
W: Spatial Data
Th: Spatial Data

June 23-26 Reducing the number of variables in a data set
Lecture 1 Space and Time Data
Tu: Reducing dimensions
W: Reducing dimensions
Th: Functional data

June 30- July 3 Models for unusual data points
M: Accommodating outliers
Tu: Extremes: statistics for rare events
W: Projects
Th: Projects

Detailed topics and data sets

Week 1

Lecture 1 *R Bootcamp* Starting and ending an R session. Basic syntax, working with vectors and arrays. Data frames and lists. Plotting concepts. Using help.

Lecture 2 *Programming in R.* Creating a function. For loops. Data objects. Using apply for subsets. Building a complex figure by simple steps

Lecture 3 *Tools to summarize and visualize data.* Summarize and visualize complex multivariate-spatial-temporal data with uncertainty Using boxplots, histograms, image, apply, split, to find patterns in data. Case study Comparing regional climate model simulations to observed data. R: (apply, hist, bplot, bplot.xy)

Lecture 4 *EDA tools continued* Case study: Comparing predictions of future climate change.

Week2

Lecture 1 Fitting distributions Multivariate summaries: mean vectors, covariance matrices, multivariate normal distribution Graphics for multivariate data, scatterplot matrices, 3-D scatterplots. R:(pairs, rgl)

Lecture 2 Regression Least-squares and its properties. Using the lm function, formula specification, interpreting output, residuals, prediction. R: (lsfit, lm)

Lecture 3 Regression How computations are done. An introduction to linear algebra in R for regression. Regression with categorical variables and variable selection. R: (chol, qr, lm, glm, spam)

Lecture 4 Fitting curves and surfaces to data. Using R functions for smoothers and splines. Connection with regression and the principle of cross-validation for evaluating predictions. R: (KernSmooth, gam, sreg, Tps)

Week 3

Lecture 1 Times Series. Correlation functions and dependence. Finding regular patterns and trends. Autoregressive models for dpendence.

Lecture 2 Times Series. Making forecasts.

Lecture 3 Spatial Data Covariance functions, Kriging and spatial predictions. R: (Krig, mKrig)

Lecture 4 Spatial Data Working with large spatial data sets. Variograms for regular locations. R: Creating covariance functions and conditional simulation.

Week 4

Lecture 1 Data with space and time components. Autoregressive models and Markov Random fields. R: (LatticeKrig, Rmpi)

Lecture 2 Reducing dimensions Principal components (EOFs)

Lecture 3 Reducing dimensions Lasso and related methods.

Lecture 4 Functional data

Week 5

Lecture 1 Extremes Generalized extreme value and the generalized Pareto distributions. R: (extRemes)

Lecture 2 Projects

Lecture 3 Projects

Lecture 4 Projects