Mangalore University

Department of Statistics
Mangalagangothri, Karnataka – 574199

Global Initiative on Academic Network (GIAN)
A course for PG Students, Researches, Faculty members, Officers/Policy Markers on

Title of the Course: Statistical Analysis of Big Data Using R

December 03-08, 2018

by

Dr. George Ostrouchov
Senior Data Scientist, Computer Science and Mathematics Division, Oak Ridge National Laboratory &
Joint Faculty Professor, Business Analytics and Statistics Department, University of Tennessee, USA
&
Prof.Ismail B
Professor, Department of Statistics, Mangalore University, Mangalagangothri.

Preamble: Global Initiative of Academic Networks(GIAN) is an initiative of higher education supported by the Government of India by inviting the talent pool of scientists and entrepreneurs internationally to encourage their engagement with the institutes of higher education in India to augment the country's existing academic resources, accelerate the pace of quality reform and elevate India's scientific technological capacity to global excellence.

Overview of the Course:

Big data analytics is the process of examining large data sets to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. Big data will gain importance not only for enterprise business but also for nations and their citizens. Big data will be the future of business and would also play key role in providing huge benefits.

The goal of this course is to introduce its participants to the interdisciplinary and emerging field of data science. The R language and its several thousands of packages for statistical analysis of diverse data are now gaining popularity across many disciplines. This workshop will cover topics beginning with the basics of R programming, through modelling and predictive data analysis with Big Data. We will cover strategies for computing with R on multi core processors and even on large distributed cluster computers for truly large data.

R and its packages are free and open source. R is able to run across Windows, Mac, and UNIX architectures. It is capable of efficient utilization of multi core platforms as well as large distributed architectures with thousands of nodes, providing unmatched data science scalability among technical computing languages.

2.0 Objectives / Learning Outcomes

The course will cover the main idea about the big data, learn data loading techniques, perform data analytics using R. Also, there will be some exploration of practical use of the methods in real applications. The course will cover main ideas from a conceptual perspective as well as investigating aspects of the underlying theory and computation. At the end of the course, the participants will be trained so as to work on the real life project on Big data analytics.

3.0 Teaching Faculty with Allotment of Lectures and Workshops

Dr.George Ostrouchov, Senior Scientist, Oak Ridge National Laboratory & Joint Faculty Professor, University of Tennessee

Lectures:	10 lecture hours
Practicals:	08 hours

4.0 Course Details

4.1 Tentative Duration: **December 03-08, 2018**

4.2 Tentative Lecture & Schedule

(Lecture hrs 10:00 am – 12:30 noon; Practicals 14:00 hrs 16:00 hrs)

Statistical Analysis of Big Data Using R

Unit 1: Introduction to R and Its Basic Data Analysis Packages

(0 hr 15 mts) Class Introductions and Logistics

(1 hr 00 mts) Introduction to Big Data and Why R

- What is big data?
- Why use R to analyze big data?
- What is High Performance Computing?

(1 hr 15 mts) Introduction to Data Analysis in R Studio

- R Studio
- Help system
- Finding and installing packages
- Basic programming
- Examples involving Distributions, Descriptive statistics, and Basic inference

(1 hr 15 mts) Selected Topics in Data Analysis

- Regression Models
- ANOVA and Multiple Testing
- Generalized Linear Models

(1 hr 15 mts) Selected Topics in Data Analysis (continued)

- Clustering and Classification
- Regression Trees and Random Forest
- Principal Components Analysis

Unit 2: R Programming and Advanced Data Analysis

(1 hr 15 mts) Intermediate R Programming

- Numerical tools (optim, integrate, uniroot)
- Debugging
- Getting your data into R: CSV files, Binary files, SQL databases, SAS datasets, Big datasets

(1 hr 15 mts) Evaluating Variability and Uncertainty

- Boot strap
- Cross validation
- Writing simulations to study properties

(1 hr 15 mts) Bayesian Computing

- Monte Carlo integration
- Importance sampling
- Markov Chain Monte Carlo: Gibbs sampling, Metropolis-Hastings

(1 hr 15 mts) Graphics

- Traditional graphics
- ggplot2
- Using knitr to create analysis documents from R code

Unit 3: High Performance R Programming

(1 hr 15 mts) Speeding up your R code

- Modular code practice
- Profiling
- Vectorizing
- Inserting C++ code into R (Rcpp)

(1 hr 15 mts) Parallel Computing

- Brief introduction to parallel hardware and software
- Using the parallel package in R

(1 hr 15 mts) Distributed Parallel Computing with pbdR

- pbdMPI - SPMD and Managing communication
- Parallel Bootstrap
- Parallel Crossvalidation
- Parallel random Forest

Examination: December 08, 2018, 114. 00 hrs -15.00 hrs

5.0 Who can attend?

- Those who are working in Statistics and wish to extend their knowledge of the tools available for Big Data analysis
- Those working in a scientific or other application area where quantitative modeling and analysis are essential and have some familiarity with Standard Statistical Methods.
- Post graduate students of Statistics and Computer Science, Research Scholars, Faculty members, Scientists from research institutes and Personnel from industry with knowledge on Statistics.

5.1 Course Fees(Rs):

1	Participants from abroad	$500
2	Faculty/Scientists from Academic Institutions/Universities/ Research Institutions	Rs.3000,
3	Participants from Private Sectors /Industries	Rs.5000/-
4	Research Scholars	Rs.1000/-
5	PG students	Rs.750/-

Foreign Faculty: **George Ostrouchov**

Department: **Business Analytics & Statistics**
Title: Joint Faculty Professor
Education : 1984-Iowa State University, Ph.D. Statistics
1978-University of Waterloo, B.Math. Statistics (Honours, Co-op)
Contact : E: go@tennessee.edu

Dr. George Ostrouchov is a Senior Data Scientist in the Computer Science and Mathematics Division of the Oak Ridge National Laboratory, USA, and Joint Faculty Professor in the Business Analytics and Statistics Department at the University of Tennessee. He holds a Ph.D. in Statistics from Iowa State University (USA), obtained after a B.Math. in Statistics from the University of Waterloo (Canada). He is Fellow of the American Statistical Association and Fellow of the American Association for the Advancement of Science, recognized for his contributions to statistical computing, particularly to enable parallel computation on big data with statistical software. He currently leads the pbdR.org project, a set of highly scalable R packages for distributed and parallel computing and profiling in data science. His recent pbdR presentation at the 2017 Intel HPC Developer Conference won the People’s Choice Award in the High Productivity Languages track.

Host Faculty: **Prof.Ismail B**

Prof. Ismail B, Professor of Statistics and is currently, the Chairman of the Department of PG Studies and Research in Statistics, Mangalore University, Mangalagangothri, Karnataka

Email: prof.ismailb@gmail.com. He was Common wealth Fellow 2000 and worked as Honorary Senior Research Fellow, Department of Statistics, University of Glasgow, Glasgow, Scotland U.K.,2000-2001.

His research interest is on Nonparametric regression**.** Worked on modeling discontinuous phenomenon. In the field of regression analysis contribution has been made on detecting discontinuities, estimation of jump size and testing for discontinuities. Developed improved methods for estimation and testing change points in regression curves and surfaces.

- Improved methods in terms of computation time and accuracy are developed for estimating discontinuities and jump size in regression curves and surfaces using non-parametric regression.
- Estimation of error variance in non-parametric regression model is developed.
- A global test for detecting the presence of discontinuities in a regression function is invented.
- Developed improved method of estimation in Econometric model with stochastic restrictions and multicollinearity problem.
- Modeling Financial time series to improve forecasting accuracy.
- More than 20 research papers have been published in the area of non-parametric regression in leading journals such as journal of Non-parametric Statistics, Journal of Statistics and Computing, Springer, Netherlands, Statistical Methods, Journal of Indian Society of Agricultural Statistics, Journal of Communications in Statistics Theory and methods, Marcel Dekker, Journal of Data Sciences and Journal of Wavelet theory.