2nd RUFORUM TRIENNIAL CONFERENCE

TRAINING CONCEPT NOTE

SCIENTIFIC DATA MANAGEMENT FOR POST-GRADUATE STUDENTS AND EARLY CAREER RESEARCHERS USING R PROGRAMMING LANGUAGE

Date: 5th-9th August 2024 **Time:** 8:30-16:00 South African Time (SAT)

Venue: International University of Management, Namibia (TBC), Windhoek, Namibia

Registration Link: XXXXXX (include questions for pre-training assessment)

Contact:
Dr. Runyararo Rukarwa (r.rukarwa@ruforum.org)

BACKGROUND

Scientific data management is a set of practices, procedures, techniques, and tools that enable scientists, in particular PhD, MSc students, and young researchers, to manage data rationally to produce quality scientific reports. In the same way, its good mastery enhances their ability to meaningfully engage in conducting quality research by developing appropriate research proposals, designing of studies, collecting, and analyzing data. The purpose is to preserve data in a consistent, accessible, secure, and uncluttered form. Without data management, data analysis will produce incorrect results and lead to biased decision making. Indeed, it is currently observed that with the technological development, and the concern to have more precise and accurate results, experiments or surveys are performed on a large scale sometimes leading to complex designs, and to subsequent messy data. Figuring out how to handle data resulting from such experiments/surveys takes time, and getting appropriate assistance is difficult. The students also do not know how to effectively analyse the data using appropriate statistical software, interpret the results, and communicate properly with the target audience.

Given these shortcomings, the Regional Universities Forum for Capacity Building in Agriculture (RUFORUM) has organised a training to provide post-graduate students and early career researchers with the skills and abilities to conduct their research effectively and efficiently. The content of the

modules in this training focuses on the techniques of processing, exploration, inferential analysis, and modelling of qualitative and quantitative data most encountered in the literature to solve real-life problems. Students and young researchers will also be exposed to the R programming language for data management, analysis, and reporting. R is a free, open-source statistical programming language primarily used by statisticians and data miners. According to the PYPL index in 2020, R is popularly ranked 7th among scholarly users worldwide (https://darly.solutions/the-most-popular-programming-languages-in-2021/). It is a very flexible in performing tasks and anyone interested in data analysis and any user can quickly learn it, whether they are a data scientist or not. R is freely accessible and meets the needs of students and researchers in developing countries, most of whom cannot afford to spend large sums of money each year to renew licenses for commercially available statistical software.

In data science with applications in many areas of life such as agriculture, biology, health, environment, economics, insurance, bioinformatics, statistics, etc., R plays a crucial role in pre-processing, exploration, and visualization of massive, varied, structured or unstructured data, prediction, classification, and clustering thanks to its multiple and diversified functionalities. A user of R can modify the various functions of R and create his own packages through execution of codes. It can also be used to develop amazing web applications, has a large support community through bootcamps and R meetings. It is well maintained, and R updates are always available. R is released under the GNU General Public License and there are no restrictions on its use. As for RStudio, it is an integrated development environment (IDE), which also supports statistical computing, graphics, and statistical models. An RStudio user can manipulate data and may be able to store used R commands for future use. The environment also provides an R markdown feature that allows work to be converted into different formats such as Word, PDF, PowerPoint, HTML, etc. This is a boon for academics, as scholarly papers can be written directly in the R markdown environment and then published in a manuscript.

Given its vast advantages for its community of graduate students, mid-career researchers and senior faculty, RUFORUM has organized a practical training in scientific data management with practical applications in R. The Scientific data management training will deliver on the following modules.

i) ***Introduction to R programming language and data management.*** This module will cover basics of R use, including downloading R packages, arithmetic operations with R, assigning variables, basic data types in R, importing and exporting datasets, defining a directory, saving datasets in R and in excel sheets.

ii) ***Exploratory analysis and reporting with R Markdown.*** This module will cover types of data, descriptives statistics, and data visualization in R such as histograms, scatter plots, box plots, and bar charts. It will also cover introduction to the production of a statistical report using R codes and generation of comments with R Markdown whose documents are fully reproducible and obtainable in PDF, HTML, Word and PPT formats among others.

iii) ***Univariate parametric and non-parametric statistical inference with R.*** This module will cover univariate statistical tests (comparison of proportions and means, correlation, t-tests, analysis of variance, chi-square tests, correlations tests such as pearson, kendall, etc.). Reporting of results, interpreting statistical analyses, discussing limitations, and drawing valid conclusions will also be covered.

iv) ***Linear regressions in R (simple and multiple).*** This module will cover principle and scripts of simple and multiple regression models, estimation, and significance test of coefficients in the model, validation of the model, analysis of variance tables, coefficient of determination (R^2) and adjusted R^2, Influential values, graphical and statistical analysis of residuals, correlations between explanatory variables, use of the model in forecasting etc.

OBJECTIVES

The overall aim of the statistics data management training is to provide postgraduate students and early career researchers with technical skills to conduct scientific research. Specifically, the participants will be able to;

a) Download R, RStudio, and R packages and install them on their computers.
b) Import data sets in different formats such as Excel, Txt, CSV into R and export R data sets to Excel, Txt and CSV.
c) Manipulate their field data in R before performing any data analysis using R's interactive command prompts.
d) Describe data sets with descriptive statistics parameters (central tendency, dispersion, and shape parameters) calculated with the R programming language.
e) Use R Markdown to store various R commands, write R scripts, add comments, which can be converted to pdf, HTML, Word document and Power Point.
f) Visualize data with the R programming language to produce quality graphs for use in manuscripts and technical reports.

g) Use the R programming language to analyze data with inferential statistical tools such as univariate parametric and non-parametric statistical testing, categorical data analysis techniques and linear regression models.

APPROACH

The delivery mode will be a mix of interactive theoretical sessions and practical work. The approach will be participatory, with students expected to be active learners and to engage in intensive and critical self-directed learning. The assignments will be designed to train and test critical thinking skills. Real data sets provided by the facilitators or obtained from the students before the course starts will be used in the examples and exercises. The training will be delivered face-to-face (physically). Each participant will need a laptop, a good internet connection, and a dataset. The daily programme will be divided into sessions providing an overview of the topics, followed by practical computer exercises, and a discussion of the statistical results. First, the basic principles, followed by examples of syntax in the R language will be presented. The participants will analyze their data using the techniques already introduced in their daily work. Discussions on the interpretation and presentation of results will take place each day during the plenary sessions. Participants will evaluate the modules daily and deficiencies will be corrected immediately. An overall evaluation of the modules will be carried out at the end of the training.

PARTICIPANTS

Participants will include postgraduate Students and early career researchers.

ORGANIZERS

The event is organized by RUFORUM, with support from the Government of Namibia and RUFORUM member universities in Namibia.

TRAINING FACILITATORS

The following facilitators will deliver the training at UNAM:

1. Prof. Susan Tumwebaze (susantumwebaze@gmail.com);
2. IUM Team (TBC)

LEARNING OUTCOMES

Day one: Participants will use R programming language to manage data and writing the code for the different statistical methods.

After day 1 of training in R programming we expect participants to have understood the use of R in data management.

Day two: Participants will have attained knowledge on how to describe data sets using numerical summaries, graphical representations, and prepare reports with R Markdown.

Day three and day four: Participants will have attained knowledge on the complete overview of statistical methods for data analysis. They will know which method to use depending on the data available and the objectives to be achieved.

Day five: Participants will have acquired methodological and practical knowledge of linear regression methods to obtain an explanatory analysis of a phenomenon, to confirm hypotheses, to take decisions or to make forecasts. A synthesis of the 4 modules will be made followed by a discussion with all participants on these models. This discussion will be related to the tasks they will have been given or to their own experiences in data management and analysis.

DETAILED TRAINING PROGRAMME (To be developed by the facilitators)